A pipeline for analysis of allele specific expression reveals salinity-dependent regulation in Nile tilapia

Aurora Campo (ayla.bcn@gmail.com)
Agricultural Research Organization
Moran Gershoni
Agricultural Research Organization
Adi Doron-Faigenboim
Agricultural Research Organization
Dietmar Kültz
University of California, Davis
Avner Cnaani
Agricultural Research Organization

Article

Keywords:

Posted Date: January 22nd, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3735302/v1

Additional Declarations: No competing interests reported.

A pipeline for analysis of allele specific expression reveals salinity-dependent regulation in Nile tilapia

Aurora Campo[1,*], Moran Gershoni[1], Adi Doron-Faigenboim[2], Dietmar Kültz[3], Avner Cnaani[1,*]

1. Institute of Animal Science, Agricultural Research Organization, Rishon LeZion, Israel
2. Institute of Plant Science, Agricultural Research Organization, Rishon LeZion, Israel
3. Department of Animal Sciences, University of California, Davis, CA, USA

Correspondence:
Aurora Campo, Department of Poultry and Aquaculture, Institute of Animal Sciences, Agricultural Research Organization, Volcani Center, Rishon LeZion, Israel. E-mail: ayla.bcn@gmail.com
Avner Cnaani, Department of Poultry and Aquaculture, Institute of Animal Sciences, Agricultural Research Organization, Volcani Center, Rishon LeZion, Israel. E-mail: avnerc@agri.gov.il

Abstract

Species living in changing environments require the acclimatization of individual organisms, which may be significantly influenced by allele specific expression (ASE). Data from RNA-seq experiments can be used to identify and quantify the expressed alleles. However, conventional allele matching to the reference genome creates a mapping bias towards the reference allele that prevents a reliable estimation of the allele counts. We developed a pipeline that allows identification and unbiased quantification of the alleles corresponding to an RNA-seq dataset, without any previous knowledge of the haplotype. To achieve the unbiased mapping, we generate two pseudogenomes by substituting the alternative alleles on the reference genome. The SNPs are further called against each

pseudogenome, providing two SNP data-sets that are averaged for calculation of the allele depth to be merged in a final SNP calling file. The pipeline presented here can calculate ASE in non-model organisms and can be applied to previous RNA-seq data-sets for expanding studies in gene expression regulation.

Introduction

High-throughput RNA-seq is a common technique in many fields of research, providing information on differentially expressed genes (DEGs) in response to different conditions or experimental factors [1]. Gene expression is quantified based on read counts that correspond to a particular gene. These reads contain single-nucleotide polymorphism (SNP) sites, providing information on the presence of different alleles. Heterozygous sites typically have equal transcription of both alleles, resulting in an allele specific expression (ASE) levels of 0.5 for each. A difference in the expression ratio between the two alleles is referred to as allelic imbalance (AI). This variation in expression patterns highlights the significance of heterozygosity and genetic variation in adaptation and acclimatization [2–5].

It is possible to quantify ASE by calculating the frequencies of expressed alleles [6–8]. To quantify alleles' expression, it is necessary to identify heterozygous sites in a gene and estimate the counts for each site. However, the most significant challenge of this measurement arises when mapping two different alleles to a reference genome, with only one being identical to the reference. The mismatch between the non-identical allele and the reference genome will discard some reads, leading to an overestimation of the identical allele [9]. This mapping bias affects the accuracy of the allele counts in gene expression experiments and creates difficulty in finding the regulatory effects of ASE [10,11].

To resolve this bias, prior knowledge of the sample haplotype is required, from either DNA-sequencing or available genotype data and reference haplotypes like HapMap [12] or SNP panels [13–15]. New approaches use multiple reference genomes for a broader view of SNPs in a population [16]. However, this is not applicable in RNA-seq experiments where genomic DNA is not sequenced.

We have developed a pipeline to call SNPs and analyze ASE, using RNA-seq data from experiments characterizing DEGs under different environmental conditions (Figure 1). Our pipeline enables quantification of ASE on coding sequences of the genome in non-model organisms without prior genotypic knowledge. We overcome the mapping bias by creating two pseudogenomes based on SNPs from two samples from different experimental groups. A final SNP data-set is generated by averaging the counts of each polymorphic site obtained from each pseudogenome, which can then be subjected to statistical tests to assess the relationship between allelic expression and environmental or physiological factors. Finally, the genomic locations of ASE SNPs can be correlated with other gene expression data, such as DEGs and methylation sites, to complement the results.

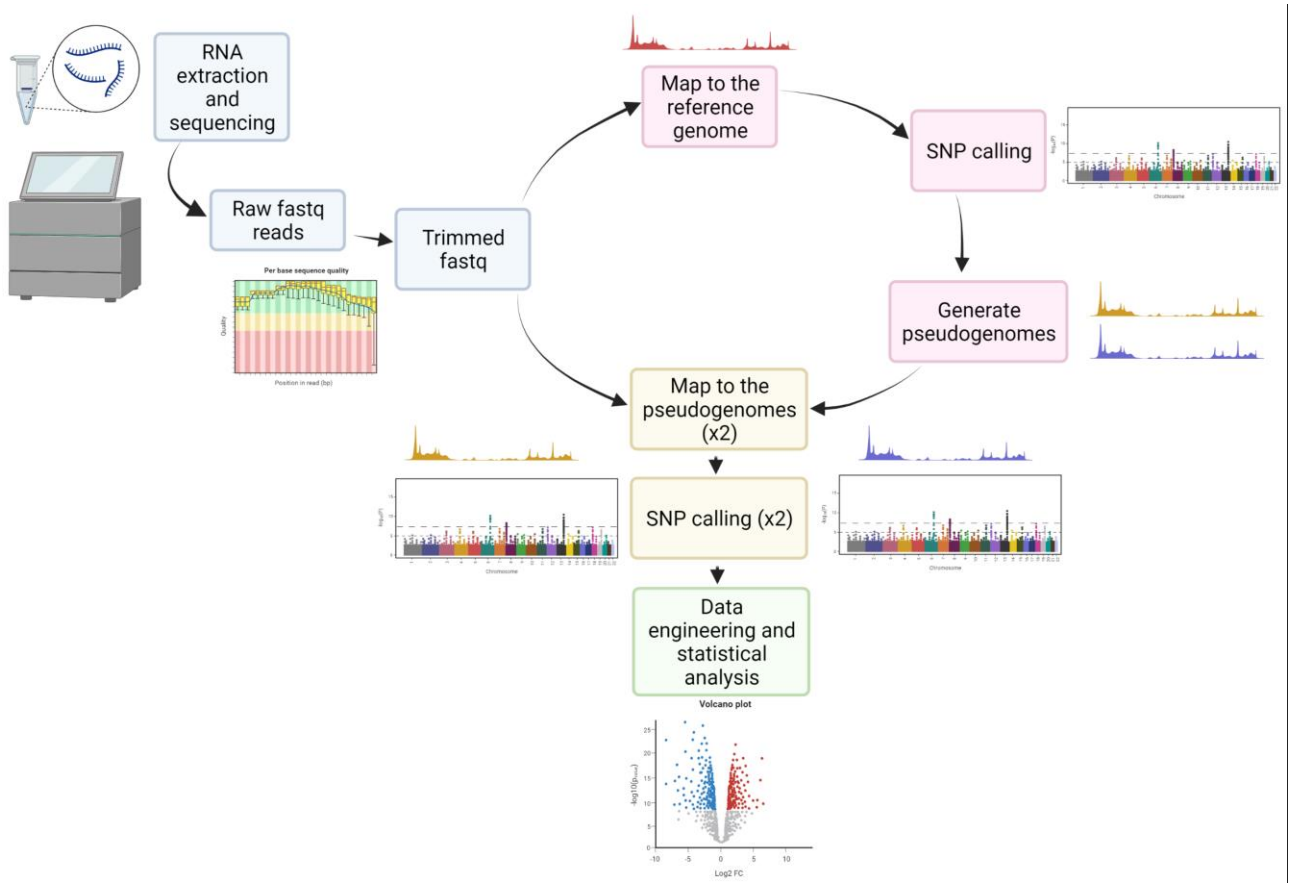

Figure 1: Schema representing the key steps of the pipeline. The RNA is extracted from the experimental samples and sequenced for obtaining of the *fastq* files. These files are trimmed and after quality filters are mapped to the reference genome for a first SNP calling. The biallelic variant sites obtained in this first call are then used for the creation of two pseudogenomes. The *fastq* files are then mapped twice, one to each pseudogenome, and the SNP call is also performed twice. The resulting variant call files are then submitted to home scripts for the merging and averaging of the allele depths. The pipeline is developed in Snakemake and the scripts were submitted to GitHub.

In this work, we applied our pipeline to study the effect of high salinity on Nile tilapia (*Oreochromis niloticus*), a freshwater fish. We performed a quantification with minimal bias of bi-allelic sites and statistical analysis of SNPs in ASE in the gills and kidney of tilapia in freshwater and brackish water environments.

Results

Data engineering and statistical analysis

SNPs retrieved by the pipeline were filtered and ASE levels were retrieved using a chi-square test for statistical analysis. We identified 817 SNPs that were heterozygous in all 12 fish and found 63 to 611 SNPs with allele-specific expression, depending on the test (Table 1, Supplementary 1-4). The allele frequency analysis of the identified SNPs within each experimental group was visualized through a heat-map (Figure 2). A consistent allele frequency pattern was observed for both gill and kidney tissues, with notable discrepancies between the two tissues. The groups exposed to salty water exhibited a heightened prevalence of allelic imbalances, either favoring the reference or alternative alleles (Figure 2).

Test description	Test number	SNPs in ASE	Genes containing ASE	DEGs with SNPs in ASE
Gills and kidney in fresh water	1	474	324	250
Gills and kidney in salty water	2	611	406	326
Gills in fresh and salty water	3	63	55	17
Kidney in fresh and salty water	4	347	240	0

Table 1: Number of SNPs obtained in each test after applying the pipeline. Namely: test 1 gills and kidney in freshwater, test 2 gills and kidney in salty water, test 3 gills in fresh and salty water, test 4 kidney in fresh and salty water. Chi-square test on the ASE-levels data for each individual. Five individuals in each group, n=10 for each test, $p<0.05$.

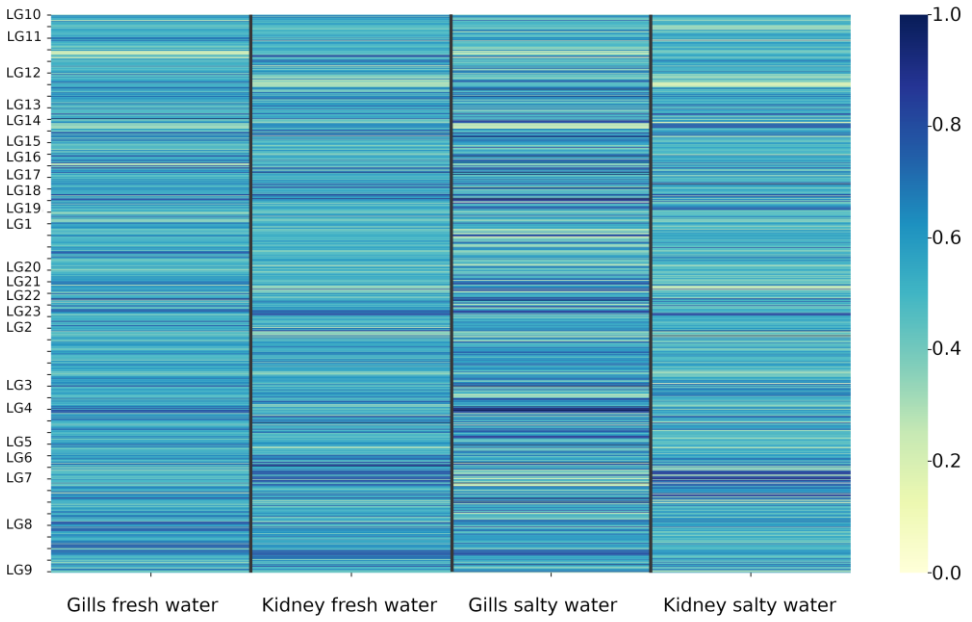

Figure 2: Heat-map showing allelic imbalance between the experimental groups. The Y-axis shows the labels of the chormosomes in wich the SNPs are located. The sequence of SNPs by chromosome follows a sequencial order from the beginning of the chromosome till the end according to the annotation provided for the reference genome.

ASE was observed in response to different salinity treatments in the gills and kidney tissues. The Chi-square tests showed 25 common SNPs displaying ASE independent of tissue or salinity treatment, and 444 SNPs in common for the tissue when comparing gills and kidney in freshwater and salty water (tests 1 and 2). There were 33 common SNPs for the salinity challenge in the gills and kidney that were related to the tissue differences (Table 1, Figure 3).

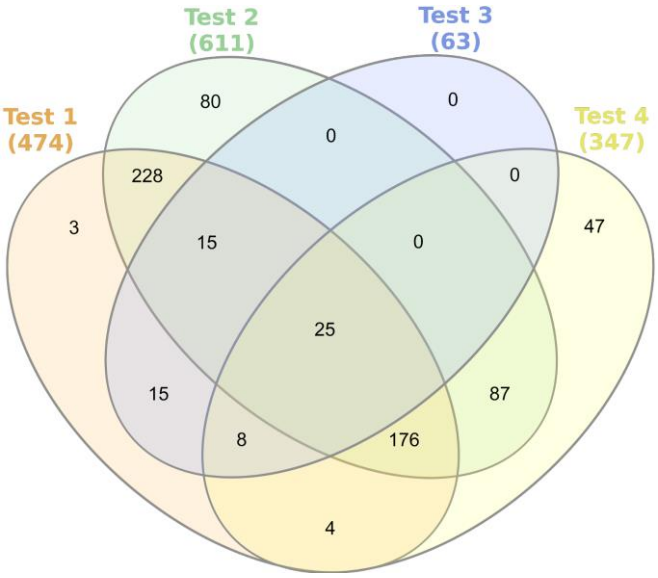

Figure 3: Venn diagram illustrating the common ASE SNPs for each test. Absolute values of SNPs.

Validation of the pseudogenomes method for eliminating mapping bias in allele counts

Levene's and Anderson-Darling tests confirmed that no equal variances and no normality were achieved on the experimental groups, respectively. Therefore, a Welch's T-test was performed in order to validate the elimination of the mapping bias by comparing abundance distribution of the ASE-levels on the SNPs retrieved with the classical method, after mapping towards the reference genome, and on the same SNPs retrieved by the new pipeline using pseudogenomes for mapping. The results indicate a significant difference between the mapping to the reference genome and the use of the pseudogenomes proposed in this pipeline ($p < 0.05$, Table 2). The visualization of the test indicates that the normally distributed data corresponds to the SNPs called on the pseudogenomes, while the SNPs called from the reference genome show certain bias on the allele frequency towards the reference allele (Figure 4).

Group	T-statistic	P-value
GF	-2.32	0.0206
GS	-3.56	0.0004
KF	-2.89	0.0004
KS	-3.37	0.0008

Table 2: Statistics and p-value of each Welch t-test. The test was performed on the abundance of the ASE-levels called against the reference genome and called against the pseudogenome, n= 744 SNPs. One test was performed in each experimental group: gills fresh water, gills salty water, kidney fresh water, kidney salty water.

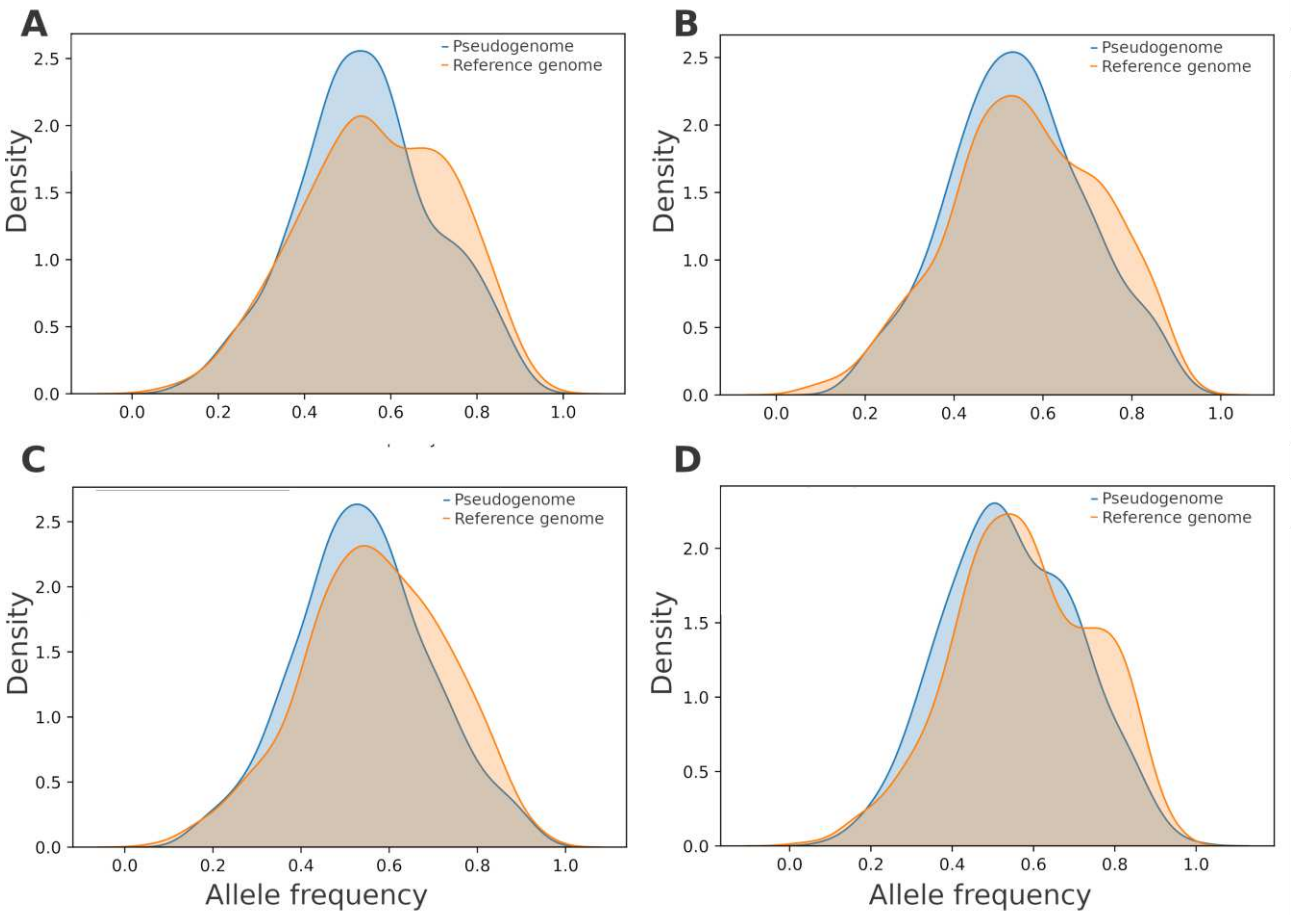

Figure 4: Student's t-test on each experimental group. The test was performed on the abundance of the ASE-levels called against the reference genome and called against the pseudogenome, n=744 SNPs. One test was performed in each experimental group.

Validation of the SNP sites by sequencing / re-sequencing

In order to test the quality of the SNP calling we compared the single nucleotide variants (SNVs) on the coding areas found on the DNA with those retrieved from RNA of the same individual. Our

findings indicate that 89.1% of SNVs, including both synonymous and non-synonymous sites, were accurately identified when calling from RNA from pseudogenomes, without using the DNA data as reference. Therefore, the error rate was of a 10.9% for SNVs in the coding areas of the genome, corresponding to a mismatch between the SNVs called from the RNA and those called from the DNA. In order to determine the source of the error, we randomly chose 20 SNVs sites among the 10.9% attributed to a mismatch between RNA and DNA SNV sites. We visualized these sites included in the *bam* files with IGV software over the DNA of the individual whose transcriptome was analyzed. A different genotype for the SNVs in the DNA than the one called by the Haplotype caller tool was found in 9 sites out of 20 mismatches. In these sites, the RNA genotypes were correctly called. A different genotype for the SNVs in the RNA than the one called by the Haplotype Caller tool was found in 11 sites out of 20 mismatches. Among these cases, 7 sites had allele counts below 5, being the low counts a possible source of error. Therefore, the partial error for identifying SNVs on the RNA call may be a fraction the total error.

Functional analysis and classification of the ASE SNPs

The function described by GO analysis of the genes containing ASE SNPs was determined for each test separately (Supplementary 5, 6, 7 and 8). The ASE variants after tissue differences conserved a similar proportion of functions both in fresh and salty water fish (Supplementary 5 for test 1 and Supplementary 6 for test 2, Table 1). Both tissues showed binding and catalytic activity as molecular functions that included ASE SNPs (tests 1 and 2, Table 1, Supplementary 5 and 6). On the other hand, some of the ASE SNPs had different gene function within the kidney and within the gills. The gills showed SNPs in ASE corresponding to response to stimulus as well as ribonucleotide binding with and without adenyl group (Supplementary 7, test 3). The kidney exposed to salinity had ASE SNPs related to developmental process and anatomical structure development differently than in the gills (Supplementary 8, test 4).

The chromosomal regions of interest for significant ASE SNPs are illustrated in the Manhattan plot (Figure 5). The ASE SNPs were distributed along all the major linkage groups of the genome for all the tests and no genomic region was found for a major abundance of ASE SNPs after exposure to salinity in gills nor kidney.

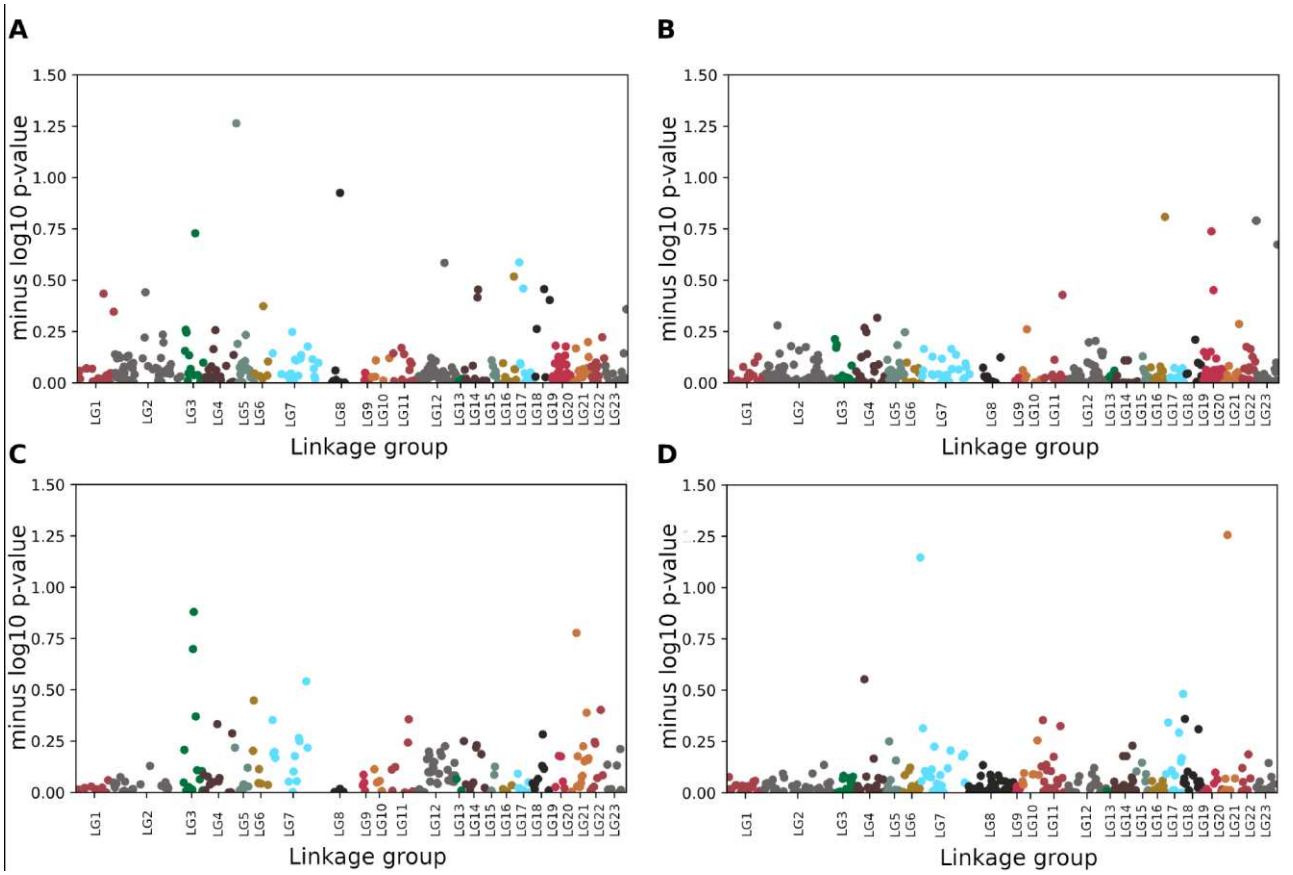

Figure 5: Manhattan plot generated with the ASE SNPs found in each test. The X-axis represents the genomic positions of the SNPs, and the Y-axis represents the significance of the associations based on the p-values obtained from the chi-square test ($p<0.05$).

The significant ASE SNPs were classified according to their function and compared for each test using a stacked bar chart (Figure 6). Notably, a substantial number of these synonymous SNPs were identified in each experimental trial, save for the gill test conducted in both fresh and saline aquatic environments (test 3), which exhibited a notably lower total SNP count. In contrast, the count of non-synonymous SNPs remained consistently below 50 across all tests. Specifically, a limited subset, fewer than 5 SNPs within each test, manifested frame-shift mutations, while instances of non-frameshift deletions and insertions were also observed. One SNP remained unassigned and was categorized as unknown variant (Figure 6).

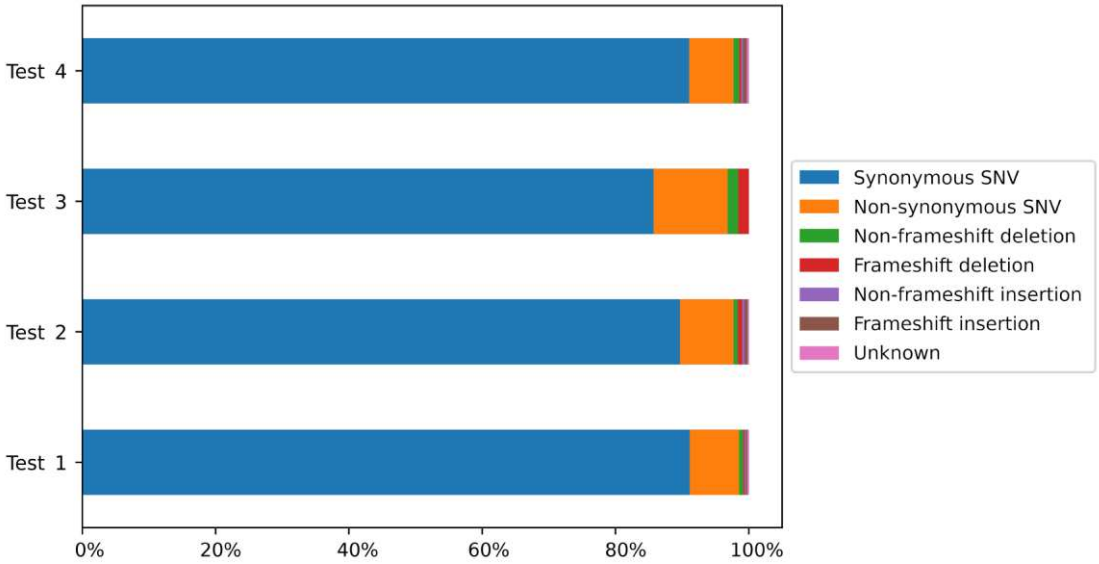

Figure 6: Classification of the ASE SNPs by the type predicted from the coordinates as set in the annotation. Test 1 gills and kidney fresh water. Test 2 Gills and kidney in salty water. Test 3 Gills in fresh and salty water. Test 4 Kidney in fresh and salty water. Chi-square test, $p<0.05$.

The analysis of differentially expressed genes revealed that 250 SNPs in ASE were also found to correspond to differentially expressed genes (ASE-DEGs) when comparing gills and kidney tissues in freshwater (test 1). Additionally, there were 326 ASE-DEGs between gills and kidney tissues in saline water (test 2) (Table 1, Supplementary 9 and 10, respectively). In the context of salinity comparison, 17 ASE-DEGs were identified in the gills (test 3) (Table 1, Supplementary 11). It is noteworthy that no ASE-DEGs associated with differentially expressed genes were detected in the kidney tissue (test 4).

Discussion

Two methods for SNP calling were compared in the present study. The first method was the commonly used, which includes the mapping of the reads to the reference genome before the SNP calling. These variants were substituted into the reference genome thus creating a pseudogenome. In the second method tested, we created two pseudogenomes corresponding to the SNPs called in two samples from different experimental groups. By doing so, we included in the pseudogenome variants expressed under two different conditions of the study thus targeting the counts of alternative alleles. The RNA data was further mapped to the pseudogenomes and the allele counts were averaged in order to get an unbiased quantification of the SNPs expression that cannot be offered by the mapping to the reference genome.

The use of pseudogenomes based on known SNPs [13,14] or by using several genomes from different heterozygous individuals [16], has been previously validated. Previous strategies used for removing mapping bias required prior knowledge of genotypes [13,17–22], elimination of sites showing bias after simulation [23–26], SNPs previously informed in a panel [27–29], or direct use of a variant-aware alignment [30–32]. However, to the best of our knowledge, this is the first approach of pseudogenomes created based on variant sites found exclusively in the RNA and without knowledge of the genotype in the DNA. The pipeline developed in this study does not require this previous knowledge. Instead, it detected the sites expressed in the individuals in the experiment. This detection allowed for the SNP quantification in those organisms whose RNA was sequenced after exposure of the individuals to different experimental conditions. For species lacking a reference genome, it is also possible to map the RNAseq data against two transcriptomes belonging to two samples of the experiment. Therefore, the new analysis pipeline is applicable to a broad range of non-model organisms.

Our pipeline followed the GATK best practices recommendations [33] and provided unbiased SNPs, from which 63 are associated to the salinity challenge in gills and 347 in the kidney. The exploration of ASE-SNPs in the context of diverse environmental conditions remains relatively limited in existing literature. Previous studies have predominantly focused on examining expression of quantitative trait loci (eQTLs) influenced by SNPs within specific regulatory regions [34,35]. Interestingly, Knowles et al. [36] developed a generalized linear model tool for analysing genome x environment interactions for ASE, known as EAGLE. This tool, although applicable to specific model organisms, such as the human liver, identified 442 ASE SNPs responsive to diverse molecular stimuli [36]. Intriguingly, the number of ASE SNPs uncovered by our pipeline in a non-model organism context falls below the range of results achieved by tools designed exclusively for model organisms. This discrepancy can be attributed to the pipeline's focus on coding sequences of the RNA.

Validation

We performed a validation of the genotype on SNP sites identified by our pipeline from RNA towards the DNAseq data of 8 individuals of tilapia exposed to freshwater. In the context of our investigation, the outcomes yield compelling insights into the accurate identification of single nucleotide variations of coding exons (SNVs), encompassing both synonymous and non-synonymous positions. The successful identification of 89.1% of these SNVs underscores the efficacy of our approach, especially considering that this error rate also includes mistakes from the SNP call on the DNA data. Moreover, a majority of the sites showing errors after the SNP call on RNA are associated with read counts below 5. As a result, implementing a stringent filter to include only counts above 5 will consistently

mitigate the error rate. Nevertheless, the potential of overlooking crucial and informative sites cautions against adopting this approach.

In summary, our study sheds light on the precision and limitations of SNV identification, with a notable success rate of ~90%. These findings contribute to a deeper understanding of the intricacies involved in SNV identification and emphasize the need for continued advancements in this field.

SNPs in ASE and regulation of gene expression

The pipeline deployed for SNP identification within the coding regions of RNA and subsequent quantification of ASE has provided valuable insights into the genetic responses of Nile tilapia under varying salinity conditions. Our approach yielded a cohort of 817 SNPs that were consistently heterozygous across all 12 fish samples. Subsequently, the investigation revealed a range of 63 to 611 SNPs exhibiting ASE, with the specific count varying according to the experimental test conducted. Notably, the analyses unveiled pronounced ASE patterns in response to diverse salinity treatments across gills and kidney tissues. The Chi-square tests conducted revealed 25 SNPs with ASE that remained consistent regardless of the tissue type or the salinity treatment, thereby indicating a degree of independence from these factors.

ASE between tissues

A substantial subset of 444 SNPs exhibited common ASE between gills and kidney tissues when comparing responses to fresh and salty water (tests 1 and 2). This shared ASE across tissues and salinity treatments suggests an underlying genetic mechanism that potentially transcends tissue specificity, evoking different allelic expression responses in the tissues independently of the salinity. In previous studies, ASE was mostly detected when comparing the expression in different tissues, as shown in cattle [37]. Also in cattle, Guillocheau et al. [17] found that 13% of the total expressed genes in muscle had SNPs in ASE associated with phenotypic traits and potentially causative of *cis*-regulation. Allelic imbalance was also described in these studies, reporting that at minimum 89% of the total SNPs were imbalanced in at least one tissue out of 18 studied [37]. Allelic imbalance was also common between 19 muscle samples of the Limousine cattle breed [17]. It has been suggested that the phenomenon of tissue-specific regulation of allele expression and the regulation of ASE may be driven by tissue-specific enhancers or by post-transcriptional differences as found in the mouse allelome [38].

In summary, our findings highlight the presence of ASE in tilapia, akin to other species, revealing a shared phenomenon of inter-tissue variation. The existence of ASE suggests the operation of complex regulatory mechanisms of expression that transcend species boundaries. To shed light on the

evolution of such a mechanism, it is imperative to explore its regulatory pathways, identifying commonalities and differences across organisms.

Salinity challenge in tilapia

Additionally, our investigation identified 33 SNPs exhibiting common ASE in response to the salinity challenge across both gill and kidney tissues. The common patterns on the GO expression for both tissues facing a salinity challenge corresponded to molecule and ion binding (Supplementary 5, 6, 7 and 8). This shared response suggests a coordinated attempt to counteract oxidative stress and uphold cellular viability in a systemic approach to achieve oxidative homeostasis between tissues and under the salinity challenge by allelic expression regulation.

Many different reasons can explain this variable expression of the alleles after salinity challenge. Gene imprinting caused by environmental factors, silencing the maternal or paternal allele, is one case [39]. The spectrum of silencing ranges from monoallelic expression (MAE), where one allele is completely silenced, to imbalances of varying degrees. Also cis-acting mutations may alter regulation for just one allele through a change to promoter/enhancer regions (transcription factor binding sites) [40], or even through 3' UTR mutations that affect mRNA stability or microRNA binding [41].

In light of these mechanisms, the gills exhibited ASE in genes associated with the ribonucleotide binding with and without adenyl group, upon salinity exposure (Supplementary 7). Earlier transcriptomic and proteomic analyses of these data indicated that there is a response in the gills to salinity by differential expression of genes related to epithelial turnover [42,43]. These previous results are consistent with our current analysis where ASE SNPs were associated to DEGs in the gills but not in the kidney (Table 1). However, the total SNP count exhibited a remarkable reduction when compared to the kidneys (test 3 and 4, Table 1). Altogether may indicate the differential expression found in gills to cope with the salinity challenge may be regulated partially by ASE, thus driving the epithelium turnover.

Meanwhile, the kidneys showed SNPs in ASE related to developmental process and anatomical structure development after salinity exposure (test 4, Supplementary 8). However, no differentially expressed genes correspond to ASE. These results align with proteomic analysis of Nile tilapia kidneys, revealing higher translational modifications in the kidney as a response to the salinity exposure in contrast with few differentially expressed genes [44]. The mechanisms in the kidney to cope with the salinity challenge were related to proteomic changes reflected already by the SNPs in ASE from our study.

A growing body of research in teleosts has unveiled ASE patterns in response to various environmental factors. In the case of turbot (*Scophthalmus maximus*), SNP studies discovered the sex determination patterns of ASE [45]. Similarly, SNP markers in Atlantic salmon (*Salmo salar*) were

found to enhance DHA synthesis [46], while eQTL were associated with resistance to lice in Atlantic salmon [47] and broad scale suppression of gene expression was detected in triploid medaka (*Oryzias latipes*) [48]. Our analysis stands in alignment with this emerging trajectory of ASE findings in teleosts facing diverse environmental challenges. Our results suggest that the salinity, as an environmental factor, may challenge each tissue in a different manner. The gill's response is characterized by a higher number of DEGs including ASE SNPs. In contrast, the kidney's response involves a higher frequency of ASE SNPs, correlating with proteomic changes previously determined [44]. These results collectively emphasize the role of salinity as a pivotal environmental factor influencing allelic expression dynamics.

In conclusion, the ASE SNPs found within the context of a salinity challenge has uncovered a rich landscape of genetic responses in Nile tilapia gill and kidney tissues. Two distinct strategies employed by Nile tilapia to confront the salinity challenge through ASE. The first strategy entails a systemic response, characterized by ASE patterns intricately linked to protein binding and oxidation-reduction processes. This widespread approach reflects a concerted effort to adapt and manage the salinity stress across multiple tissues. Conversely, the second approach is marked by nuanced differentiation. In the gills, a concerted effort to address the challenge is evident through a higher prevalence of ASE related to DEGs. This tailored response is particularly notable given the reduced number of DEGs observed in this tissue. In contrast, the kidney employs a proteomic approach, where ASE patterns exhibit a broader scope, encompassing SNPs unrelated to DEGs. This distinct strategy indicates a sophisticated mechanism by which the kidney leverages allelic expression to navigate the salinity stress, potentially involving regulatory mechanisms beyond traditional gene expression pathways.

Type of SNPs in ASE

The significant SNPs in ASE, underwent additional analysis using a stacked bar chart, presented in Figure 6. The prominent SNPs identified for specific allelic expression were predominantly single nucleotide variants. Synonymous SNVs, which preserve amino acid composition within translated proteins, significantly outnumbered non-synonymous SNPs by approximately 8 to 13 times across all tests. Synonymous polymorphisms have potential impact on messenger RNA splicing, stability, structure, and protein folding [49]. These alterations can substantially affect protein function, modulate cellular responses to therapeutic targets, and frequently elucidate distinct patient reactions to specific medications [49], which becomes particularly relevant following an environmental challenge as presented in this study.

In contrast, non-synonymous SNPs consistently remained below 50 across all tests, implying a level of coding region conservation. Comparative investigations into these SNV types have revealed that diversity among non-synonymous SNVs is notably lower than among synonymous substitutions [50].

This discrepancy is attributed to natural selection's influence on non-synonymous SNPs [51], highlighting how environmental pressures influence the prevalence and distribution of different SNV types. Moreover, the identified frameshift and non-frameshift mutations amplify the scope of genetic variability and its potential functional implications. The presence of an unclassified SNP serves as a reminder of the complexities of genetic annotation and the ongoing pursuit to decipher the functional significance of genetic variation. Overall, the paragraph contributes to our understanding of how genetic variation, influenced by natural selection and environmental challenges, plays a pivotal role in shaping the evolutionary trajectories of organisms.

Regions of the genome affected by different allele expression

The pipeline targets the regions of the genome that may be affected by the different expression of the alleles in two different approaches: with a heat-map on the allelic imbalance (Figure 2) and a Manhattan plot on the ASE SNPs (Figure 5).

The visualization of allele frequency analysis for the identified SNPs within each experimental group was executed using a heat-map according to a distribution on the genome, as shown in Figure 2. Gill and kidney tissues exhibited very different allele frequency patterns, thus illustrating the major differences in allele expression depending on the tissue, as previously described [37]. Striking a consistent pattern, gills in fresh and salty water exhibited parallel allele frequency trends, yet discernible disparities were evident between the two environment types. A similar effect is visible in kidney from fresh and from salty water. Notably, exposure to saline water triggered pronounced shifts in allelic balance within the respective groups.

The identification and visualization of significant allelic expression ASE SNPs through the lens of chi-square tests across chromosomal regions were achieved using a Manhattan plot (Figure 5). No distinct genomic region emerged as an exclusive hotspot for an abundance of ASE SNPs in response to salinity exposure, neither in gills nor kidney tissues. This lack of clustering suggests a much more dispersed genetic response to salinity stress as previously indicated by the heat-map, with significant ASE SNPs dispersed across the genome rather than being concentrated in specific regions. Focusing solely on ASE SNPs from the entire pool of RNA-derived SNPs could potentially limit our ability to identify specific genomic regions responsible for regulating allelic expression. However, this approach effectively targets SNPs that undergo altered expression in response to the experimental factor being investigated – in this instance, the salinity challenge. The results of our analysis show that while there are patterns of allelic imbalance throughout the genome, those do not correspond with exposure to salinity.

In conclusion, the combined insights from the heat-map and Manhattan plot analyses presented in the pipeline offer a comprehensive understanding of the genetic dynamics underlying allelic expression

changes in response to environmental shifts. The heat-map underscores the tissue-specific allelic imbalances on the experimental groups, while the Manhattan plot showcases the distributed nature of significant ASE SNPs across the genome. This integrative approach advances our comprehension of the genetic mechanisms governing responses to environmental challenges, shedding light on the intricate regulatory networks operating within the examined organisms.

The pipeline presented in this study has proven to be a robust and versatile tool for SNP quantification, enabling unbiased assessment of SNPs in ASE resulting from fish exposures to an environmental challenge. This methodology operates without requiring prior genotype knowledge, making it suitable for model and non-model organisms, regardless of strain or the availability of a reference genome. This approach opens avenues for reanalyzing data from differentially expressed gene (DEG) experiments to uncover gene regulation and shifts in biochemical networks driven by specifically expressed alleles. Furthermore, the SNP coordinates generated by this pipeline can be seamlessly integrated with other sources of transcript data, including methylome information. This pipeline, based on Snakemake and integrating GATK best practices, tailored to the unique context of non-model organisms and has facilitated a deeper understanding of how genetic variation influences gene expression responses under different conditions. By extending the pipeline's scope to include analysis of non-coding regions, we envision a compelling prospect: the utilization of our pipeline may significantly advance the study of regulatory regions in non-model organisms. Such an expansion could potentially enhance our understanding of the intricate genetic responses these organisms exhibit in the face of environmental challenges.
The findings of this study, coupled, therefore, with insights from comparative research, underscore the intricate nature of allelic expression dynamics, tissue-specificity, and regulatory mechanisms, inviting further exploration into the epigenomic dimensions of ASE regulation in species like tilapia.

In conclusion, our developed pipeline not only reveals unbiased allele-specific expression in response to environmental challenges, as demonstrated in the context of salinity alterations in Nile tilapia, but it also prioritizes user convenience. By integrating various tools and virtual environments within the Snakemake framework, our pipeline streamlines the analytical process, empowering researchers to explore and interpret their data efficiently. Furthermore, the graphical representation of results adds a layer of visual comprehension, enhancing the accessibility and impact of our findings. This holistic approach underscores our commitment to advancing the field of allele-specific expression analysis while prioritizing usability and clarity for researchers in non-model organism studies.

Material and methods

Ethical statement

This study was approved by the Agricultural Research Organization Committee for Ethics in Experimental Animal Use, and was carried out in compliance with the current laws governing biological research in Israel (Approval number: IL-715/17). The study is reported in accordance with the ARRIVE guidelines (https://arriveguidelines.org).

Origin, processing and sequencing of samples

The sequences used in this study were from an experiment performed in Nile tilapia fish previously described by Root et al.[43,44]. In summary, 12 male Nile tilapia were randomly allocated into two 600-liter freshwater tanks and allowed to acclimate for two weeks. One group was subjected to a gradual increase in salinity of 5 ppt per day until reaching a final salinity of 25 ppt, at which point gill and kidney samples were collected after 24 hours. In addition, eight fish were individually placed in 40 liter freshwater tanks and allowed to acclimate for two weeks before sampling gills for RNAseq and fin for DNAseq.

The samples were preserved in RNAsave (Biological Industries, Beit-Haemek, Israel) and stored at -20°C until extraction, following a 24-hour incubation period at 4°C. The mRNA of the first 12 individuals' samples was extracted using TRIzol reagent (Thermo Fisher Scientific), and purified using the TURBO DNA-free kit (Invitrogen) to eliminate DNA contamination. The quality of the total mRNA samples was assessed at the Israel National Center for Personalized Medicine (INCPM) at the Weizmann Institute of Science (Rehovot, Israel) using the TapeStation Agilent 2200 system before library preparation and sequencing on an Illumina Hi-Seq 2500 device. For the additional eight individuals, the mRNA and DNA were extracted using the RNeasy mini kit and DNeasy blood and tissue kit (Qiagen, Hilden, Germany), respectively. RNA and DNA sequencing for the additional eight individuals were conducted at the INCPM using the Illumina Novaseq 600 platform, with the RNA samples sequenced to an average sequencing depth of 10 times, and the DNA samples sequenced to an average sequencing depth of 30 times, both of which included unique molecular identifier (UMI) barcoding.

Pipeline for mapping bias removal using pseudogenomes and SNP calling

We developed and utilize a Snakemake-based [52] pipeline for SNP calling and removal of mapping bias using pseudogenomes. We implemented this pipeline with and without prior knowledge of genotypes, and the entire code along with scripts is available at GitHub (https://github.com/AylaScientist/Snakemake_for_SNPs). *Fastq* files were trimmed with

Trimmomatic [53] and quality was verified with FASTQC (v0.11.8, Andrews, 2010) before mapping against the *O. niloticus* reference genome (NMBU GCF_001858045.2) using STAR (v2.7.1a, Dobin et al., 2013). SNPs were called following GATK best practices [33] and Picard tools were used (release 2.27.5). Two pseudogenomes were constructed from a *vcf* file following the protocol by Johan Zicola (https://github.com/johanzi/make_pseudogenome, MIT license) for downstream analysis. Two final *vcf* files joining the SNPs from all samples were annotated using ANNOVAR [56]. Allele depth and genotypes were collected into a table (VariantsToTable, GATK).

Data engineering and statistical analysis

After the pipeline produced two data-sets, each containing alternative SNPs for all sites identified in one of the pseudogenomes, Python v3.7.3 scripts developed in-house and part of the Snakemake pipeline were used to merge and filter the data-sets for multiallelic sites using Pandas and Numpy. The counts of each reference and alternative polymorphic site were then averaged for each sample. SNPs with a depth of one allele less than 3 and those with a total allele depth less than 10 were removed to avoid homozygosis. As the genotype is unknown, only the SNPs shared by all individuals with both alleles expressed were retained in the final data-set of consensus SNPs.

Validation of the pseudogenomes method for eliminating mapping bias in allele counts

To validate the mitigation of mapping bias, a comparison was conducted by assessing allele frequencies of ASE-levels in two scenarios: SNPs obtained through the conventional method involving mapping to the reference genome, and SNPs retrieved via the novel pipeline, which employed pseudogenomes for mapping. After studying the normality and the equal variances with Anderson-Darling and Levene's tests respectively, Welch's t-test was performed on the common SNPs called by each method.

Validation of the SNP sites by sequencing / re-sequencing

We used a pipeline to identify SNPs from both DNA and RNA, and conducted a qualitative analysis to detect false positives by comparing SNPs identified from RNA with those present in DNA. We selected SNPs that had biallelic expression and were exonic single nucleotide variants (SNVs), excluding indels and intronic sites. The mismatches were attributed to the error from SNP calling from RNA information. We further used IGV software [57] to analyze 20 SNVs for false negatives.

Functional analysis and classification of the ASE SNPs

Significant ASE SNPs for each treatment were analyzed for gene ontology (GO) functions. DEG analysis was performed with the DESeq package [58] in R (v 3.6.3, Development Core Team, 2013) for

484 salinity effects. In order to find regulatory pathways, the SNPs in ASE were contrasted with the
485 significant DEGs.
486

487 Data availability

488 The sequencing data was submitted to SRA under the bioproject PRJNA669315.
489

490 Code Availability

491 The snakemake pipeline was submitted to the GitHub
492 https://github.com/AylaScientist/snakemake_for_SNPs.

493

494 Acknowledgements

495 This research was supported by grant 2016611 from the US-Israel Binational Science Foundation
496 BSF and the US-Israel Binational Agricultural Research and Development Fund (BARD) Grant (IS-
497 5358-21).
498
499

500 References

501 1. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. RNA-seq: An assessment
502 of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18**,
503 1509–1517 (2008).
504 2. Barrett, R. D. H. & Schluter, D. Adaptation from standing genetic variation. *Trends Ecol. Evol.*
505 **23**, 38–44 (2008).
506 3. Lande, R. & Shannon, S. The Role of Genetic Variation in Adaptation and Population
507 Persistence in a Changing Environment Author (s): Russell Lande and Susan Shannon
508 Published by : Society for the Study of Evolution Stable URL :
509 http://www.jstor.org/stable/2410812 Accessed : 07. *Evolution (N. Y).* **50**, 434–437 (1996).
510 4. Hermisson, J. & Pennings, P. S. Soft sweeps: Molecular population genetics of adaptation from
511 standing genetic variation. *Genetics* **169**, 2335–2352 (2005).
512 5. Bernatchez, L. On the maintenance of genetic variation and adaptation to environmental
513 change: considerations from population genomics in fishes. *J. Fish Biol.* **89**, 2519–2556

(2016).

6. Garber, M., Grabherr, M. G., Guttman, M. & Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods* **8**, 469–477 (2011).
7. Ozsolak, F. & Milos, P. M. RNA sequencing: Advances, challenges and opportunities. *Nat. Rev. Genet.* **12**, 87–98 (2011).
8. Trapnell, C. *et al.* Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. *Nat. Biotechnol.* **28**, 511–515 (2011).
9. Degner, J. F. *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**, 3207–3212 (2009).
10. Monsu, M. & Comin, M. Fast alignment of reads to a variation graph with application to SNP detection. *J. Integr. Bioinforma.* **18**, 1–9 (2021).
11. Zhan, S., Griswold, C. & Lukens, L. Zea mays RNA-seq estimated transcript abundances are strongly affected by read mapping bias. *BMC Genomics* **22:285**, 1–12 (2021).
12. Consortium, I. H. A haplotype map of the human genome The International HapMap Consortium. *Physiol. Genomics* **437**, 1299–320 (2003).
13. Rozowsky, J. *et al.* AlleleSeq: Analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.* **7**, 1–15 (2011).
14. Vijaya Satya, R., Zavaljevski, N. & Reifman, J. A new strategy to reduce allelic bias in RNA-Seq readmapping. *Nucleic Acids Res.* **40**, 1–9 (2012).
15. Harvey, C. T. *et al.* QuASAR: Quantitative allele-specific analysis of reads. *Bioinformatics* **31**, 1235–1242 (2015).
16. Chen, N., Solomon, B., Mun, T., Iyer, S. & Langmead, B. Reference flow : reducing reference bias using multiple population genomes. *Genome Biol.* **22**, 1–17 (2021).
17. Guillocheau, G. M. *et al.* Survey of allele specific expression in bovine muscle. *Sci. Rep.* **9**, 1–11 (2019).
18. Yuan, S. & Qin, Z. Read-mapping using personalized diploid reference genome for RNA sequencing data reduced bias for detecting allele-specific expression. *Proc. - 2012 IEEE Int. Conf. Bioinforma. Biomed. Work. BIBMW 2012* 718–724 (2012) doi:10.1109/BIBMW.2012.6470225.
19. Xin, H. *et al.* Accelerating read mapping with FastHASH. *BMC Genomics* **14**, 1–13 (2013).
20. Braasch, I. *et al.* The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nat Genet* **48**, 427–437 (2016).
21. Mayba, O. *et al.* MBASED: Allele-specific expression detection in cancer tissues and cell lines. *Genome Biol.* **15**, 1–21 (2014).

549 22. Pandey, R. V., Franssen, S. U., Futschik, A. & Schlötterer, C. Allelic imbalance metre (Allim),
550 a new tool for measuring allele-specific gene expression with RNA-seq data. *Mol. Ecol.*
551 *Resour.* **13**, 740–745 (2013).
552 23. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation
553 with RNA sequencing. *Nature* **464**, 768–772 (2010).
554 24. Panousis, N. I., Gutierrez-arcelus, M., Dermitzakis, E. T. & Lappalainen, T. Allelic mapping
555 bias in RNA-sequencing is not a major confounder in eQTL studies. *Genome Biol.* **467**, 1–8
556 (2014).
557 25. Stevenson, K. R., Coolon, J. D. & Wittkopp, P. J. Sources of bias in measures of allele-specific
558 expression derived from RNA-seq data aligned to a single reference genome. *BMC Genomics*
559 **14**, 1–13 (2013).
560 26. Hodgkinson, A., Grenier, J. C., Gbeha, E. & Awadalla, P. A haplotype-based normalization
561 technique for the analysis and detection of allele specific expression. *BMC Bioinformatics* **17**,
562 1–10 (2016).
563 27. Van De Geijn, B., Mcvicker, G., Gilad, Y. & Pritchard, J. K. WASP: Allele-specific software
564 for robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063 (2015).
565 28. Salavati, M. *et al.* Elimination of Reference Mapping Bias Reveals Robust Immune Related
566 Allele-Specific Expression in Crossbred Sheep. *Front. Genet.* **10**, 1–16 (2019).
567 29. Gutierrez-Arcelus, M. *et al.* Allele-specific expression changes dynamically during T cell
568 activation in HLA and other autoimmune loci. *Nat. Genet.* **52**, 247–253 (2020).
569 30. Hach, F. *et al.* MrsFAST-Ultra: A compact, SNP-aware mapper for high performance
570 sequencing applications. *Nucleic Acids Res.* **42**, 494–500 (2014).
571 31. Buchkovich, M. L. *et al.* Removing reference mapping biases using limited or no genotype
572 data identifies allelic differences in protein binding at disease-associated loci. *BMC Med.*
573 *Genomics* **8**, 1–15 (2015).
574 32. Miao, Z., Alvarez, M., Pajukanta, P. & Ko, A. ASElux: An ultra-fast and accurate allelic reads
575 counter. *Bioinformatics* **34**, 1313–1320 (2018).
576 33. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples.
577 *bioRxiv* 201178 (2017) doi:10.1101/201178.
578 34. Wang, J. The computer program structure for assigning individuals to populations: easy to use
579 but easier to misuse. *Mol. Ecol. Resour.* **17**, 981–990 (2017).
580 35. Zhang, F. *et al.* Genetic architecture of quantitative traits in beef cattle revealed by genome
581 wide association studies of imputed whole genome sequence variants: I: feed efficiency and
582 component traits. *BMC Genomics* **21**, 1–22 (2020).
583 36. Knowles, D. A. *et al.* Allele-specific expression reveals interactions between genetic variation

584 and environment. *Nat. Methods* **14**, 699–702 (2017).
585 37. Chamberlain, A. J. *et al.* Extensive variation between tissues in allele specific expression in an
586 outbred mammal. *BMC Genomics* **16**, 1–20 (2015).
587 38. Andergassen, D. *et al.* Mapping the mouse Allelome reveals tissue-specific regulation of allelic
588 expression. *Elife* **6**, 1–29 (2017).
589 39. Monk, D., Mackay, D. J. G., Eggermann, T., Maher, E. R. & Riccio, A. Genomic imprinting
590 disorders: lessons on how genome, epigenome and environment interact. *Nat. Rev. Genet.* **20**,
591 235–248 (2019).
592 40. Serre, D. *et al.* Differential allelic expression in the human genome: A robust approach to
593 identify genetic and epigenetic Cis-acting mechanisms regulating gene expression. *PLoS*
594 *Genet.* **4**, (2008).
595 41. Mayr, C. Regulation by 3 -Untranslated Regions. *Annu. Rev. Genet.* **51**, 171–194 (2017).
596 42. Campo, A. *et al.* Different transcriptomic architecture of the gill epithelia in Nile and
597 Mozambique tilapia after salinity challenge. *Comp. Biochem. Physiol. - Part D Genomics*
598 *Proteomics* **41**, 100927 (2022).
599 43. Root, L. *et al.* Nonlinear effects of environmental salinity on the gill transcriptome versus
600 proteome of Oreochromis niloticus modulate epithelial cell turnover. *Genomics* **113**, 3235–
601 3249 (2021).
602 44. Root, L. *et al.* A data-independent acquisition (DIA) assay library for quantitation of
603 environmental effects on the kidney proteome of Oreochromis niloticus. *Molecuar Ecol.*
604 *Resour.* (2021) doi:10.22541/au.160553713.37893872/v1.
605 45. Martínez, P. *et al.* Multiple evidences suggest sox2 as the main driver of a young and complex
606 sex determining ZW/ZZ system in turbot (Scophthalmus maximus). *bioRxiv* 834556 (2019)
607 doi:https://doi.org/10.1101/834556.
608 46. Horn, S. S., Meuwissen, T. H. E., Moghadam, H., Hillestad, B. & Sonesson, A. K. Accuracy
609 of selection for omega-3 fatty acid content in Atlantic salmon fillets. *Aquaculture* **519**, 734767
610 (2020).
611 47. Robledo, D. *et al.* Discovery and functional annotation of quantitative trait loci affecting
612 resistance to Sea lice in Atlantic salmon. *Front. Genet.* **10**, 1–10 (2019).
613 48. Garcia, T. I. *et al.* Novel method for analysis of allele specific expression in triploid Oryzias
614 latipes reveals consistent pattern of allele exclusion. *PLoS One* **9**, 100250 (2014).
615 49. Hunt, R., Sauna, Z. E., Ambudkar, S. V, Gottesman, M. M. & Kimchi-Sarfaty, C. Silent
616 (Synonymous) SNPs: Should We Care About Them? BT - Single Nucleotide Polymorphisms:
617 Methods and Protocols. in *Single Nucleotide Polymorphisms Methods and Protocols* (ed.
618 Komar, A. A.) 23–39 (Humana Press, 2009). doi:10.1007/978-1-60327-411-1_2.

50. Graur, D. & Li, W. *Fundamentals of molecular evolution*. (Sinauer Associates, Incorporated Publishers, 1997).
51. Ohta, T. Synonymous and Nonsynonymous Substitutions in Mammalian Genes and the Nearly Neutral Theory. *J. Mol. Evol.* **40**, 56–63 (1995).
52. Köster, J. & Rahmann, S. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
53. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
54. Andrews, S. FastQC: a quality control tool for high throughput sequence data. (2010).
55. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
56. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, 1–7 (2010).
57. Robinson, J. T., Thorvaldsdóttir, H., Wenger, A. M., Zehir, A. & Mesirov, J. P. Variant review with the integrative genomics viewer. *Cancer Res.* **77**, e31–e34 (2017).
58. Anders, S. *et al.* Differential expression analysis for sequence count data via mixtures of negative binomials. *Adv. Environ. Biol.* **7**, 2803–2809 (2010).
59. Development Core Team, R. R: A language and environment for statistical computing. (2013).

Author contributions

Author Aurora Campo produced the data analysis on the experimental data, and the development of the code, testing and validation of the pipeline, as well as the RNA and DNA extraction, sequencing and analysis on the dataset for validation of the pipeline. This author performed the writing of the article. Authors Moran Gershoni and Adi Doron-Faigenboim supervised some steps of the pipeline, as for example merge of databases or use of UMI. Author Dietmar Kültz supervised the experimental design and of the writing of the article. Author Avner Cnaani supervised the experimental design and the sequencing protocol of the experimental data in which the pipeline was tested, as well as supervision of the writing.

Competing interest statement

The authors declare that they have no conflict of interest related to the research, authorship or publication of the present article.

Supplementary material

Supplementary 1: Table including ASE SNPs contained in genes that are differentially expressed between gills and kidney in freshwater (test 1). Coordinates of chromosome and position, as well as Chi-square statistic, product description and GO function associated.

Supplementary 2: Table including ASE SNPs contained in genes that are differentially expressed between gills and kidney in salty water (test 2). Coordinates of chromosome and position, as well as Chi-square statistic, product description and GO function associated.

Supplementary 3: Table including ASE SNPs contained in genes that are differentially expressed between gills in fresh and salty water (test 3). Coordinates of chromosome and position, as well as Chi-square statistic, product description and GO function associated.

Supplementary 4: Table including ASE SNPs contained in genes that are differentially expressed between kidney in fresh and salty water (test 4). Coordinates of chromosome and position, as well as Chi-square statistic, product description and GO function associated.

Supplementary 5: Top 10 most frequent GO functions of the ASE SNPs found in test 1, gills in fresh water compared to kidney in fresh water. The functions are separated by Molecular Function,

Biological Process and Cellular Component. The x axis indicate the counts of SNPs in ASE corresponding to genes associated to the GO function. The intensity of the color of the bar indicates the adjusted p-value on the correspondent function associated to the set of genes in Nile tilapia.

Supplementary 6: Top 10 most frequent GO functions of the ASE SNPs found in test 2, gills in salty water compared to kidney in salty water. The functions are separated by Molecular Function, Biological Process and Cellular Component. The x axis indicate the counts of SNPs in ASE corresponding to genes associated to the GO function. The intensity of the color of the bar indicates the adjusted p-value on the correspondent function associated to the set of genes in Nile tilapia.

Supplementary 7: Top 10 most frequent functions of the ASE SNPs found in test 3, gills in fresh water compared to gills in salty water. The functions are separated by Molecular Function, Biological Process and Cellular Component. The x axis indicate the counts of SNPs in ASE corresponding to genes associated to the GO function. The intensity of the color of the bar indicates the adjusted p-value on the correspondent function associated to the set of genes in Nile tilapia.

Supplementary 8: Top 10 most frequent GO functions of the ASE SNPs found in test 4, kidney in fresh water compared to kidney in salty water. The functions are separated by Molecular Function, Biological Process and Cellular Component. The x axis indicate the counts of SNPs in ASE corresponding to genes associated to the GO function. The intensity of the color of the bar indicates the adjusted p-value on the correspondent function associated to the set of genes in Nile tilapia.

Supplementary 9: Table including the differentially expressed genes between gills and kidney in freshwater (test 1) that present ASE SNPs for the same test.

Supplementary 10: Table including the differentially expressed genes between gills and kidney in salty (test 2) that present ASE SNPs for the same test.

Supplementary 11: Table including the differentially expressed genes between gills after salinity challenge (test 3) that present ASE SNPs for the same test.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementary1Test1CHIpfdrASEdescriptionnoMAEexonic.xls
- Supplementary2Test2CHIpfdrASEdescriptionnoMAEexonic.xls
- Supplementary3Test3CHIpfdrASEdescriptionnoMAEexonic.xls
- Supplementary4Test4CHIpfdrASEdescriptionnoMAEexonic.xls
- Supplementary5GOtest1.pdf
- Supplementary6GOtest2.pdf
- Supplementary7GOtest3.pdf
- Supplementary8GOtest4.pdf
- Supplementary9ASEDEtest1tissuefreshnilenoMAEexonic.xls
- Supplementary10ASEDEtest2tissueseanilenoMAEexonic.xls
- Supplementary11ASEDEtest3gillsnoMAEexonic.xls