Identification of SNPs associated with muscle yield and quality traits using allelic-imbalance analyses of pooled RNA-Seq samples in rainbow trout

Rafet Al-Tobasei 1, Ali Ali 2, Timothy D. Leeds 3, Sixin Liu 3, Yniv Palti 3, Brett Kenney 4 and Mohamed Salem 1,2*

Abstract

Background: Coding/functional SNPs change the biological function of a gene and, therefore, could serve as “large-effect” genetic markers. In this study, we used two bioinformatics pipelines, GATK and SAMtools, for discovering coding/functional SNPs with allelic-imbalances associated with total body weight, muscle yield, muscle fat content, shear force, and whiteness. Phenotypic data were collected for approximately 500 fish, representing 98 families (5 fish/family), from a growth-selected line, and the muscle transcriptome was sequenced from 22 families with divergent phenotypes (4 low- versus 4 high-ranked families per trait).

Results: GATK detected 59,112 putative SNPs; of these SNPs, 4798 showed allelic imbalances (>2.0 as an amplification and <0.5 as loss of heterozygosity). SAMtools detected 87,066 putative SNPs; and of them, 4962 had allelic imbalances between the low- and high-ranked families. Only 1829 SNPs with allelic imbalances were common between the two datasets, indicating significant differences in algorithms. The two datasets contained 7930 non-redundant SNPs of which 4439 mapped to 1498 protein-coding genes (with 6.4% non-synonymous SNPs) and 684 mapped to 295 lncRNAs. Validation of a subset of 92 SNPs revealed 1) 86.7–93.8% success rate in calling polymorphic SNPs and 2) 95.4% consistent matching between DNA and cDNA genotypes indicating a high rate of identifying SNPs with allelic imbalances. In addition, 4.64% SNPs revealed random monoallelic expression. Genome distribution of the SNPs with allelic imbalances exhibited high density for all five traits in several chromosomes, especially chromosome 9, 20 and 28. Most of the SNP-harboring genes were assigned to important growth-related metabolic pathways.

Conclusion: These results demonstrate utility of RNA-Seq in assessing phenotype-associated allelic imbalances in pooled RNA-Seq samples. The SNPs identified in this study were included in a new SNP-Chip design (available from Affymetrix) for genomic and genetic analyses in rainbow trout.

Keywords: Rainbow trout, Fish, SNPs, Genetic markers, RNA-Seq

Background

Fish growth rate, muscle yield and fillet quality are major traits affecting profitability of aquatic food animal production. As feed cost is a major factor influencing the profitability, efficiency of growth is important and related to growth rate and muscle yield and composition. Skeletal muscle constitutes about 50–60% of the fish weight [1]. Given that growth efficiency and fillet firmness and appearance are critical for profitability and production of premium products [2], optimizing fish growth, muscle yield and fillet quality traits is a key objective in aquaculture breeding programs. Traditional phenotype-based selection is typically used to select for fast growth; however, muscle yield and quality traits are difficult to improve by conventional selection because measurement of these traits requires sacrificing the animal [2].
Genomic selection tools have been created to improve economically important traits in plants and livestock. Genetic maps, which characterize the linkage or co-inheritance patterns of genetic markers, have been developed for a wide range of species, including fish, with the aim of discovering allelic variation affecting traits; and ultimately identify DNA sequences underlying phenotypes [3, 4]. Markers have been identified by various molecular techniques, including numerous and genomewide single nucleotide polymorphisms (SNPs). In addition, recent technological developments have enabled high throughput genotyping of these SNPs rendering them useful for genome-wide association studies [5–8]. Functional SNPs are generally defined as SNPs from genome sequences that affect structure, expression or function of a gene. These sequences include coding SNPs (e.g. non-synonymous, splicing), promoter and noncoding SNPs, as well as functional elements identified from studying of genome conservation [9]. Coding/Functional/ SNPs (c/fSNPs) are especially important because they have the potential to change the function of a protein [4, 10, 11]. In addition, c/fSNP markers, because they are located within expressed genes, they are unlikely to become unlinked from their associated genes due to genetic recombination. Therefore, c/fSNPs can be useful genetic markers for detecting significant associations with phenotypes. Understanding molecular mechanisms of muscle growth and quality can help in making better selection decisions. In terrestrial livestock, several genes, genetic markers and QTLs associated with production traits, including growth, have been characterized using molecular techniques [12, 13]. In addition, marker-assisted selection has been used to enhance genetic improvement in livestock breeding programs by direct selection on genes affecting economic traits [14] and to optimize selection for quantitative traits [12, 13]. However, the genetic basis of muscle growth and quality traits is not well studied in fish [15].

Rainbow trout is the most cultivated cool and cold freshwater fish in the U.S. [16], and it is considered a model species for studies in several fields of biology, including ecology [17], pathology [18], physiology [19], toxicology [20] and carcinogenesis [21]. Several studies used RNA sequencing to identify markers in human [22, 23] and non-model species [11, 24, 25]. However, most SNP detection algorithms were developed for DNA-Seq analyses and are not optimized/tested for RNA-Seq, especially in pooled samples. The objective of this study was using RNA-Seq analyses of pooled samples to identify c/fSNP markers and develop a resource for studies of marker association with production traits in rainbow trout. First, transcriptome-wide SNP allele frequencies were correlated to phenotypic variations in fish whole body weight (WBW) and muscle yield, fat content, shear force and whiteness. Second, SNPs with allelic imbalance scores (ratios between the allelic frequencies of the high-end families and that of the low-end families) were identified. Then, a subset of the putative SNPs was validated for allelic polymorphism and tested for trait association. Finally, genes harboring SNPs with allelic imbalances were annotated to obtain insight into the potential functional effects of the SNPs.

Result and discussion
Phenotypes
SNPs were identified in fish families with divergent phenotypes in WBW, muscle yield, fat content, shear force (texture) and whiteness of the fillet. These rainbow trout were from a growth-selected line developed by the NCCCWA breeding program [26]. Briefly, this line was created through artificial selection, starting in 2004, from 7 founder strains with documented diversity and domestication history. Over five generations, the population responded to selection by 9.8–12.7% increase in WBW per generation, and rate of inbreeding averaged 0.86% per generation [26]. In this study population, which was sampled after three generations of selection (hatch year of 2010), WBW was positively correlated with muscle yield and muscle fat content ($R^2 = 0.56$ and 0.50 respectively, data not shown). Our previous reports showed that fast growth may be genetically associated with improved muscle yield, paler fillets (affected by intramuscular fat content) and firmer texture [27]. The trait heritability estimates for muscle yield, muscle weight, WBW10, WBW13, carcass weight, fat percentage, shear force and fillet color were moderate to high (0.31–0.81) [6, 27]. Those moderate to high heritability estimates imply that substantial additive genetic variation exist in the study population for growth and carcass traits.

For RNA sequencing, muscle samples were collected from 7 to 9 different full-sib families showing divergent phenotypes per trait (i.e. 3–5 high ranked families versus 3–5 low ranked families per trait). Five fish were sampled from each family. Divergent phenotypic attributes (Fig. 1) were statistically different ($P < 0.01$): WBW (1221.6 g ± 84.25 vs. 502.1 ± 28.0 g), muscle yield (50.9% ± 1.6 vs. 43.3% ± 2.3), muscle crude-fat (9.24% ± 1.2 vs. 4.77% ± 1.3), shear force (grams force/g of sample; 539.64 ± 12.3 vs. 310.01 ± 49.2), and fillet whiteness index (44.7 ± 0.8 vs. 41.23 ± 0.4) for high- vs. low-ranking groups, respectively. Means and standard deviations of these traits were calculated from the family averages.

Identification of putative SNPs
RNA pools from muscle tissues of 5 fish per family were used for RNA-Seq analyses. A total of 259,634,620 reads (100 bp single-end) were generated from 22 families at
an average of 11,801,573 reads per family. Reads were aligned against the rainbow trout genome [28] using the STAR [29] alignment tool. Percentage of reads mapped to the genome ranged from 80% to 82% per family.

A total of 204,604 putative SNPs were detected for the five traits using Haplotypecaller tool of Genome Analysis Toolkit v3.3.0 (GATK) [30], with an average of 40,920 SNPs per trait. Using the SAMtools/Popoolation software package [31, 32], a total of 304,805 putative SNPs were predicted, with an average of 60,961 SNPs per trait (Table 1). After removing redundant SNPs among all traits, we had 59,112 SNPs from GATK and 87,066 from SAMtools/Popoolation2 with 50,885 shared between the two bioinformatics pipelines (Table 1).

After identifying putative SNPs, an in-house Perl script was used to estimate allelic imbalances of the SNPs in each trait. A total of 6275 SNPs with allelic imbalances were identified from the GATK dataset at cutoff values of >2.0 as an amplification and <0.5 as loss of heterozygosity. In addition, 969 SNPs explicitly existed in only the high or low phenotypic group. After removing redundant SNPs between traits at the two cutoff values, there were 4798 unique SNPs (Table 1). Similarly, SAMtools/Popoolation2 identified 5070 SNPs with allelic imbalances at cutoff values of >2.0 as an amplification and <0.5 as loss of heterozygosity. In addition, 1450 SNPs existed in families at one of the two ends of each trait variation scale but not in the other (Table 1). There were 4962 non-redundant SNPs among the five traits that were identified with SAMtools/Popoolation2 at the two cutoff values. There were only 1829 non-redundant SNPs shared between GATK and SAMtools/Popoolation2. Differences in variant calling and filtering steps might have caused the observed differences in number of SNPs between GATK and SAMtools/Popoolation2. There were 7930 non-redundant SNPs with allelic imbalances from both methods. The results of the SNPs’ allelic imbalances should be taken with caution because we could not find a reliable statistical test associated with the ratio calls derived from the allelic imbalance calculation to report statistical significance. However, by utilizing exact allele counts instead of frequencies, we were able to assign Chi Square P-Values to most of the SNPs with

![Fig. 1 Phenotypic variations in fish families with contrasting phenotypes for five different traits; whole-body weight (\(a\)), muscle yield (\(b\)), fat content (\(c\)), shear force (\(d\)) and fillet whiteness index (\(e\)). All differences were statistically significant (\(p < 0.01\)).](image-url)
allelic imbalances. Out of the 7930 SNPs with allelic imbalances, there were 6038 SNPs with available read count for both alleles in the divergent families. Alternatively, there were 1892 with counts for only one allele. These SNPs existed in families at one of the two ends of each trait variation scale but not in the other. We performed chi-square test on the 6038 SNPs and found 5330 SNPs (83%) with $P_{\text{value}} < 0.05$ and 710 SNPs (17%) with $P_{\text{value}} > 0.05$ (Additional file 1).

For subsequent analyses, we combined SNPs from GATK and SAMtools/Popoolation2 into three different groups: 1) Non-redundant SNPs with allelic imbalances from both methods (7930 SNPs); 2) Common putative SNPs from both methods (50,885 SNPs); 3) Putative non-redundant SNPs from both methods (95,234 SNPs) (Table 1). All SNPs data are provided in Additional file 1.

**SNP validation**

A total of 92 putative SNPs including 88 SNPs from the GATK and SAMtools/Popoolation2 datasets were selected for SNP validation. Among the 92 putative SNPs, 68 SNPs showed allelic imbalances (Table 2), including 25 SNPs identified by GATK pipeline, 10 SNPs identified by SAMtools pipeline, and 33 SNPs identified by both pipelines (Table 2). Among the 92 tested SNPs, 72 (78.2%) SNPs were polymorphic, 11 (11.9%) SNPs were monomorphic and 9 failed the assay (Table 2). Failure of the Fluidigm assay can be caused by unsuccessful or non-specific primer binding to the target genomic DNA. Therefore, we cannot assume that a failed assay indicates failure of our bioinformatics pipeline to detect a SNP in the RNA sequence data, and can remove the failed SNP assays from the calculation of SNP validation rate. As 72 out of the 83 working Fluidigm SNP assays were polymorphic, we can claim 86.7% validation rate in detecting polymorphic SNPs in the overall putative SNP pool and 90% validation rate in the GATK/SAMtools shared SNPs pool. This success rate is much higher than what we previously achieved in rainbow trout using RNA-Seq (70%) and genomic reduced representation libraries (48%) [11, 33]. The improved success rate in this study is perhaps due to use of a reference genome instead of de novo assembled references used in the previous studies. In addition, a transcriptome sequence coverage of $\sim7.4X$ per fish was used compared to only $\sim0.97X$ in our previous RNA-Seq study [11]. The 90% successful SNP validation rate is comparable to that reported in diploid fish or using genomic RADs and doubled haploid fish in rainbow trout [7, 34]. In addition, a recent rainbow trout genome re-sequencing study with at least 10X genome coverage per fish had 86% successful validation rate [7]. Relatively lower success rates in SNP detection were

### Table 1

Summary of putative SNPs and SNPs showing allelic imbalances identified by SAMtools and GATK for each trait

| Trait   | No. of putative SNPs | No. of SNPs with Allelic imbalance |
|---------|----------------------|-----------------------------------|
|         | SAMtools/Popoolation2 | GATK | SAMtools/Popoolation2 | GATK |
| Fat%    | 59,032               | 38,808 | 662 | 406 | 877 | 270 |
| Shear   | 60,309               | 38,960 | 910 | 488 | 1152 | 261 |
| Muscle% | 61,117               | 42,383 | 1321 | 116 | 1507 | 76 |
| Whiteness | 64,636               | 44,460 | 1011 | 347 | 1283 | 298 |
| WBW     | 59,711               | 39,993 | 1166 | 93 | 1456 | 64 |
| Total # SNPs | 304,805              | 204,604 | 5070 | 1450 | 6275 | 969 |
| Total # SNPs non-redundant | 87,066 | 59,112 | 4962 | 4798 |
| Total Common SNPs | 50,885 | 1829 |

Allelic imbalances were calculated at $>2$ for amplification and $<0.5$ for loss of heterozygosity. SNPs explicitly existing in only the high or low phenotypic group are indicated in the table by the 0.0/1.0 ratio. * 59 SNPs were multi-allelic, showing different alleles in association with different phenotypes. ** 1 SNP was multi-allelic showing different alleles predicted by different pipelines

### Table 2

Number of putative and validated SNPs from each dataset

| SNP Group                          | Total SNPs | Polymorphic | Monomorphic | Failed assay | Success rate |
|------------------------------------|------------|-------------|-------------|--------------|--------------|
| All putative SNPs (95,234)         | 92         | 72          | 11          | 9            | 86.7%        |
| GATK/SAMTool common SNPs (50,289)  | 88         | 72          | 8           | 8            | 90.0%        |
| Total SNPs with allelic imbalance  | 68         | 55          | 7           | 6            | 88.7%        |
| GATK unique SNPs with allelic imbalance | 25   | 21          | 3           | 1            | 87.5%        |
| SAMTool unique SNPs with allelic imbalance | 10  | 4           | 2           | 4            | 66.7%        |
| GATK/SAMTool common SNPs with allelic imbalance | 33  | 30          | 2           | 1            | 93.8%        |
reported from RNA-Seq studies in rainbow trout due to
genome duplication and assembly errors in the genome/
transcriptome references [11, 35, 36]. Noteworthy and in
a separate study, we found variation in gene expression
in only 75 genes distributed between all 5 traits (data
will be published elsewhere). Therefore, differential gene
expression effects on estimating allelic imbalances were
negligible as only 75 genes distributed between all five
traits were differentially expressed between the high and
low families. Minor effects of variation in gene expres-
sion on allele frequency estimation accuracy were pre-
viously reported [37]. The SNP validation data, albeit
small, indicated that the GATK method was more suc-
cessful in calling polymorphic SNPs with allelic imbal-
ances than the SAMtools pipeline; 87.5% versus 66.7%,
respectively. However, combined GATK and SAMtools
data had a 93.8% success rate. Success rates between
SNPs with and without allelic imbalances were 88.7%
and 86.7%, respectively. Importantly and out of 72 vali-
dated SNPs, 61 (84.7%) and 58 SNPs (80.5%) were poly-
morphic in fish from two different commercially
important rainbow trout populations in the US, Trou-
lodge Inc. and Clear Springs Foods Inc., respectively.
These results suggest that the SNPs identified in this
study are also useful for other commercial rainbow trout
populations.

Assessment of mono-allelic gene expression

Out of the 72 validated polymorphic SNPs (Table 2),
there were 46 SNPs that showed potential mono-allelic
expression in cDNA in at least one fish. In other words,
the genomic DNA is heterozygous for the SNP while
cDNA is monomorphic. Thirty-three of the 35 fish
showed mono-allelic expression in at least one SNP. Out
of the aforementioned 46 SNPs, 5 SNPs were randomly
selected for validation using Sanger sequencing. All
SNPs were heterozygous at the DNA level. However,
manual investigation of the cDNA sequence chromato-
grams exhibited existence of substantial allelic imbal-
ances ranging from existence of two alleles with >2.0 X
peak height ratios between the 2 alleles at the SNP base
to a complete mono-allelic expression (a single peak).
Overall, approximately 4.64% random mono-allelic/al-
lelic imbalances existed in gene expression of rainbow
tROUT. These data are consistent with a recent study in
human stem cells showing that most allelic imbalances
did not represent ‘on/off’ events, but instead revealed
biased expression from each allele [38]. None of the 8
tested families in our study showed mono-allelic expres-
sion in all individuals specific to a given family, indicat-
ing no parental origin effect through genomic
imprinting. Likewise, the human stem cell study sug-
gested that most of the allele-biased gene expression is
not due to genomic imprinting [38]. Compared to our
estimated 4.64% mono-allelic expression, recent studies
showed 12–24% random mono-allelic expression in
mammals and 7–9% in interspecies catfish [4, 39–41].
Our mono-allelic expression assessment is based on only
72 SNPs, and hence a genome-wide assessment of
mono-allelic expression in rainbow trout warrants fur-
ther investigation.

SNP genomic/functional classification

Three sets of SNPs were considered for genomic/func-
tional classifications. For the 7930 SNPs with allelic im-
balances, 2898 (37.69%) were intergenic. Of them, 635
(8.01%) and 721 (9.09%) SNPs were located within 5Kb
upstream or downstream of protein-coding genes,
respectively. The rest of the intergenic SNPs, 1633
(20.59%) were located more than 5Kb distant to protein-
coding genes.

On the other hand, 4941 (62.31%) SNPs were genic,
including 214 (2.70%) that were located within the 5’
untranslated region (5’UTR) and 1677 (21.15%) that
were located in the 3’ untranslated region (3’UTR) of
protein coding genes. In addition, 2548 (32.13%) SNPs
were located within coding DNA sequences (CDS) and
502 (6.33%) SNPs were located within introns. Of the
CDS SNPs, 504 (6.36%) were non-synonymous; 4 of
these caused early stop codon, and 500 caused amino
acid substitution (Table 3). There were 684 (8.63%) SNPs
located within 295 IncRNAs (Table 3).

Regarding the GATK/SAMtools shared SNPs (50,885
SNPs), there were 20,356 (40.00%) intergenic SNPs. Of
these shared SNPs, 4594 (9.03%) were located within
5Kb upstream, and 5208 (10.23%) downstream of
protein-coding genes. In addition, 10,554 (20.74%) were
intergenic, more than 5Kb distant to protein-coding
genes. In contrast, 30,529 (60.00%) SNPs were genic.
And, 1389 (2.73%) of these SNPs were in the 5’UTR;
10,259 (20.16%) were in the 3’UTR, 15,178 (29.83%)
were within CDS; and 3703 (7.28%) were within introns.
Out of those within CDS SNPs, 3919 (7.70%) were
non-synonymous SNPs. Fifty of these CDS SNPs were
nonsense (causing premature stop codon), and 3869
(7.60%) were missense SNPs (Table 3).

Concerning all the putative SNPs, there were 46,901
(49.25%) intergenic SNPs. Of these, 9005 (9.46%) were
located within 5Kb upstream; and 10,245 (10.76%) were
downstream of protein-coding genes. In addition, 27,651
(29.03%) were more than 5Kb distant from protein-
coding genes. Alternatively, 48,333 (50.75%) SNPs were genic, and of these genic SNPs, 2247 (2.36%) were in the 5'UTR; 16,420 (17.24%) were in the 3'UTR; 22,616 (23.75%) were within CDS; and 7050 (7.40%) were within introns. Of the CDS SNPs, 5853 (6.15%) were non-synonymous with 79 SNPs causing early stop codons and 5774 (6.06%) causing amino acid changes (Table 3).

In these three SNP datasets, there were large percentages of intergenic (including upstream/downstream) SNPs (37–49%). Approximately 10% intergenic in addition to 30% non-coding SNPs were reported in humans from RNA-Seq data [42]. Our high percentages of intergenic SNPs may be partially explained by the incomplete annotation of protein coding genes and exons in the current version of the rainbow trout reference genome sequence [28].

Distribution and density of SNPs in the genome

Chromosome density distribution of the SNPs with allelic imbalances exhibited high density for all five traits in several chromosomes with the three highest peaks in chromosomes 9, 20 and 28 (Fig. 2a). All five traits revealed very similar pattern of distribution with a single exception; shear force exhibited a relative higher density than the other traits on chromosome 9. The similarity in density distribution between traits may be explained at least in part by the positive correlation that we observed between the phenotypes in this population. WBW and thermal growth coefficient were used as selection criteria in this population [11, 26], and we found that WBW as an independent variable has significant effects on muscle yield and fat percentage (multivariable regression analysis \( P < 0.01 \), data not shown). However, despite the similarity in SNP density distributions, most of the identified SNPs were unique to each trait. From the 7930 SNPs with allelic imbalances, only 27 were shared by all five traits, 161 were shared by four traits, 680 were shared by three traits and 1783 were shared by two traits. In agreement with our results, a recent GWAS study identified two windows with effect on fillet yield located on chromosome 9 and explaining 1.0–1.5% of genetic variance in the same fish population [6].

As can be expected, the number of SNPs with allelic imbalances per chromosome was strongly correlated with chromosome length (Fig. 2b). In general, numbered unknown chromosomes, which are longer in the current reference genome [28], had more SNPs compared to the known chromosomes (Fig. 2b). Chromosome “Unknown” (1.1 Gb of scaffolds not assigned to chromosomes) had 4086 (49.05%) SNPs (not shown in Fig. 2b). Previous genetic mapping reports showed that the growth-related SNPs/QTL are distributed over ~20 chromosomes [11, 43, 44]. Together with our data, these reports confirm the polygenic nature of growth/muscle related traits in rainbow trout.

SNP functional annotation

Functional annotation of genes harboring SNPs with allelic imbalances were performed using the Blast2GO suite [45]. The SNP-flanking sequences were searched against the NCBI nr-protein database using BLASTx; then, associated genes and Gene Ontology (GO) terms were acquired. In the biological processes category, SNP-harboring genes were associated with various cellular processes mainly involved in growth-related mechanisms, including regulation of metabolic and oxidation-
reduction processes and protein translation (Fig. 3). In the molecular function category, SNP-containing genes were associated with binding metal ions, ATP, nucleic acid, and actin. In addition, a significant number of the genes were associated with transferase, motor, oxidoreductase, and structural molecule activities (Fig. 3). In the cellular component category, many of the genes exhibited association with the cytoplasmic compartment, membranes, myosin complex, and extracellular region compartment (Fig. 3). Genes with similar GO associated terms were previously reported to be involved in rainbow trout muscle growth and quality [11, 19, 43, 46–48]. Additionally, KEGG pathway mapping was used to assign enzyme function to the SNP-containing transcripts [49]. Searching transcripts against the KEGG database yielded 1043 transcripts (13.15%) with significant KEGG hits to 632 KEGG Orthologies (KOs) belonging to different pathways (Table 4). Most of the transcripts were assigned to growth-related metabolic pathways. There were 275 transcripts (182 KOs) related to metabolism. Under this category, sequences matching energy metabolism (88 transcripts, 57 KOs) appeared on the top of the list, with 52 transcripts (37 KOs) assigned to oxidative phosphorylation. Sequences matching carbohydrate metabolism occupied the second place (77 transcripts, 43 KOs) and were further classified into glycolysis/gluconeogenesis (39 transcripts, 18 KOs), citrate cycle (19 transcripts, 14 KOs) and pyruvate metabolism (16 transcripts, 10 enzymes). The next metabolic subcategories in the metabolic list were amino acid metabolism (56 transcripts, 41 KOs), lipid metabolism (27 transcripts, 22 KOs), and cofactors and vitamins metabolism (14 transcripts, 11 KOs). These preliminary SNP functional annotations are in agreement with previous reports that showed strong association between 1) mutations and altered expression of glycolytic and oxidative phosphorylation enzymes and 2) rainbow trout growth and muscle degeneration [11, 19, 43, 46, 47]. In addition, 176 KEGG annotated sequences were assigned to the genetic information processing category.
(112 KOs) that included translation (105 sequences, 69 KOs), folding, sorting and degradation (62 sequences, 38 KOs), and transcription (9 sequences, 5 KOs) (Table 4). A significant number of the SNP-harboring genes matched ribosomal (68 sequences, 48 KOs) and RNA-transport proteins (22 sequences, 12 KOs). Previously, we showed that the atrophying muscle and muscle from fast versus slow growing rainbow trout had differentially expressed genes involved in RNA processing, protein synthesis, posttranslational modification, and intracellular protein trafficking [19, 43, 46].

Moreover, 166 sequences (99 KOs) were classified by KEGG mapping into the environmental information processing category; these sequences were further assigned to signal transduction (147 sequences, 87 KOs) and RNA-processing pathways (22 sequences, 12 KOs). Previously, we showed that the atrophying muscle and muscle from fast versus slow growing rainbow trout had differentially expressed genes involved in RNA processing, protein synthesis, posttranslational modification, and intracellular protein trafficking [19, 43, 46].

Fig. 3 Gene Ontology (GO) assignment of the genes harboring SNPs with allelic imbalances in families with contrasting growth and muscle phenotypes. Genes were assigned GO terms according to their biological processes (a), molecular functions (b) and cellular components (c).

(112 KOs) that included translation (105 sequences, 69 KOs), folding, sorting and degradation (62 sequences, 38 KOs), and transcription (9 sequences, 5 KOs) (Table 4). A significant number of the SNP-harboring genes matched ribosomal (68 sequences, 48 KOs) and RNA-transport proteins (22 sequences, 12 KOs). Previously, we showed that the atrophying muscle and muscle from fast versus slow growing rainbow trout had differentially expressed genes involved in RNA processing, protein synthesis, posttranslational modification, and intracellular protein trafficking [19, 43, 46].

Moreover, 166 sequences (99 KOs) were classified by KEGG mapping into the environmental information processing category; these sequences were further assigned to signal transduction (147 sequences, 87 KOs) and signaling and interaction molecules (19 sequences, 12 KOs) (Table 4). The PI3K-Akt signaling, Calcium signaling, MAPK signaling, and cGMP-PKG signaling pathways had the largest numbers of hits: 21, 18, 18, and 16 KOs, respectively. Previous studies indicated involvement of MAPK and Calcium signaling in fish/muscle growth [46, 50].

Furthermore, the cellular processes category contained 152 KEGG-annotated sequences matching 85 KOs, which were further classified into cellular community (54 transcripts, 27 KOs), transport and catabolism (42 transcripts, 24 KOs), and cell growth and death (36 transcripts, 22 KOs) (Table 4). In the organismal systems category, the most significant subcategories were endocrine (105 transcripts, 53 KOs), circulatory (49 transcripts, 30 KOs), immune (44 transcripts, 28 KOs), and digestive systems (32 transcripts, 16 KOs). Recently, a GWAS study using the same fish population identified a small number of genes involved in muscle development explaining ~1.0% of the total genetic variance of the muscle yield and growth rate [6].

Distributions of KEGG matches were generally similar among all five traits. Albeit, we noticed an increased number of hits related to fillet whiteness compared to other traits, for carbohydrate metabolism (47 transcripts, 28 KOs) and amino acid metabolism (32 transcripts, 26 KOs) (Table 4). Similarly, there was a noticeable increase in numbers of hits in whiteness for PI3K-Akt signaling, focal adhesion, gap junction and regulation of actin cytoskeleton (Table 4). Regulation of focal adhesion and actin cytoskeleton were associated with development of pale, soft, and exudative (PSE) meat in turkey [51]. In addition, the muscle yield trait exhibited an increased number of transcripts for energy metabolism, with 28
| KEGG categories                          | Total (all traits) | WBW | Muscle % | Fat % | Schear | Witness |
|------------------------------------------|--------------------|-----|----------|-------|--------|---------|
| **KEGG categories**                     | No. of sequences (%) | No. of KOs | No. of sequences (%) | No. of KOs | No. of sequences (%) | No. of KOs | No. of sequences (%) | No. of KOs | No. of sequences (%) | No. of KOs |
| Metabolism                               | 275                | 182 | 133      | 99    | 130    | 96      | 102     | 77    | 125    | 92    | 142    | 105    |
| Carbohydrate Metabolism                  | 7 (28.00)          | 43  | 42 (31.58) | 28    | 39 (3000) | 27     | 26 (2549) | 19    | 39 (31.20) | 25    | 47 (33.10) | 28    |
| Glycollysis / Gluconeogenesis            | 39                | 18  | 19       | 10    | 19      | 14      | 13      | 7     | 21     | 11    | 20     | 13     |
| Citrate cycle (TCA cycle)                | 19                | 14  | 8        | 8     | 9       | 7       | 7       | 6     | 10     | 9     | 9      | 7      |
| Pyruvate metabolism                      | 16                | 10  | 6        | 4     | 7       | 7       | 5       | 4     | 10     | 8     | 6      | 5      |
| Pentose phosphate pathway                | 13                | 5   | 8        | 5     | 7       | 5       | 7       | 3     | 3      | 2     | 3      | 3      |
| Energy Metabolism                        | 88 (32.00)        | 57  | 42 (31.58) | 32    | 45 (3462) | 29     | 37 (3627) | 26    | 30 (2400) | 21    | 41 (2887) | 32    |
| Oxidative phosphorylation                | 57                | 43  | 42 (31.58) | 28    | 39 (3000) | 27     | 26 (2549) | 19    | 39 (31.20) | 25    | 47 (33.10) | 28    |
| Amino Acid Metabolism                    | 56 (20.36)        | 41  | 26 (19.55) | 21    | 24 (1846) | 20     | 2 (2059)  | 16    | 28 (2240) | 22    | 32 (2254) | 26    |
| Lipid Metabolism                         | 27 (9.82)         | 22  | 14 (10.53) | 11    | 11 (846)  | 11     | 7 (686)   | 6     | 16 (1280) | 14    | 13 (9.15) | 11    |
| Fatty acid degradation                   | 15                | 13  | 8        | 6     | 9       | 9       | 4       | 3     | 11     | 10    | 8      | 7      |
| Metabolism Cofactors and Vitamins        | 14 (5.09)         | 11  | 4 (3.01)  | 4     | 6 (4.62)  | 4       | 5 (490)   | 5     | 5 (400)  | 5     | 4 (282)  | 4      |
| Nucleotide Metabolism                    | 13 (4.73)         | 8   | 5 (3.76)  | 3     | 5 (385)   | 5       | 6 (588)   | 5     | 7 (560)  | 5     | 5 (352)  | 4      |
| Genetic Information Processing           | 176               | 112 | 69       | 50    | 79      | 59      | 50      | 40    | 83     | 59    | 74     | 55     |
| Translation                              | 105 (5966)        | 69  | 36 (52.19) | 31    | 48 (6076) | 39     | 30 (6000) | 27    | 45 (5422) | 35    | 47 (6351) | 40    |
| Ribosome                                 | 68                | 48  | 25       | 23    | 32      | 26      | 21      | 19    | 33     | 27    | 32     | 28     |
| RNA transport                            | 22                | 12  | 9        | 6     | 11      | 9       | 7       | 6     | 8      | 4     | 7      | 6      |
| Folding, Sorting and Degradation         | 62 (35.23)        | 38  | 28 (40.58) | 17    | 24 (3038) | 16     | 19 (3800) | 12    | 30 (3614) | 19    | 23 (3108) | 13    |
| Protein processing in endoplasmic reticulum | 23              | 14  | 11       | 7     | 7       | 4       | 7       | 5     | 14     | 9     | 7      | 4      |
| RNA Degradation                          | 16                | 5   | 11       | 4     | 10      | 5       | 8       | 4     | 9      | 3     | 11     | 5      |
| Proteasome                               | 12                | 11  | 4        | 4     | 5       | 5       | 1       | 1     | 4      | 4     | 4      | 3      |
| Ubiquitin mediated proteolysis           | 9                 | 7   | 2        | 2     | 2       | 2       | 1       | 1     | 3      | 3     | 1      | 1      |
| Transcription                            | 9 (5.11)          | 5   | 5 (7.25)  | 2     | 7 (8.86)  | 4       | 1 (200)  | 1     | 8 (964)  | 5     | 4 (541)  | 2      |
| Spliceosome                              | 9                 | 5   | 5        | 2     | 7       | 4       | 1       | 1     | 8      | 5     | 4      | 2      |
| Environmental Information Processing     | 166               | 99  | 70       | 45    | 88      | 61      | 62      | 45    | 76     | 55    | 87     | 58     |
| Signal Transduction                      | 147 (8855)        | 87  | 62 (88.57) | 39    | 79 (8977) | 53     | 56 (9032) | 41    | 69 (9079) | 50    | 74 (8506) | 50    |
| PI3K-Akt signaling pathway               | 35                | 21  | 12       | 8     | 13      | 10      | 12      | 8     | 13     | 10    | 24     | 16     |
| Calcium signaling pathway                | 36                | 18  | 16       | 9     | 13      | 8       | 14      | 10    | 16     | 12    | 16     | 11     |
Table 4  KEGG biochemical mapping of the genes harboring SNPs with allelic imbalances in fish families showing contrasting growth and muscle phenotypes (Continued)

| Signaling Pathway                        | 26 | 18 | 10 | 6 | 17 | 12 | 7 | 6 | 12 | 9 | 10 | 8 |
|------------------------------------------|----|----|----|---|----|----|---|---|----|---|----|---|
| MAPK signaling pathway                   | 21 | 12 | 10 | 5 | 14 | 8 | 9 | 3 | 10 | 6 | 12 | 7 |
| cGMP-PKG signaling pathway               | 18 | 12 | 6  | 5 | 4  | 4 | 7 | 6 | 8  | 7 | 7  | 6 |
| AMPK signaling pathway                    | 11 | 9  | 4  | 2 | 7  | 4 | 3 | 3 | 9  | 6 | 8  | 4 |
| cAMP signaling pathway                   | 13 | 7  | 2  | 2 | 7  | 5 | 6 | 6 | 5  | 5 | 5  | 4 |
| HIF-1 signaling pathway                  | 7  | 6  | 3  | 3 | 2  | 2 | 3 | 3 | 1  | 1 | 3  | 3 |
| FoxO signaling pathway                   | 5  | 5  | 1  | 1 | 2  | 0 | 2 | 0 | 1  | 1 | 3  | 3 |
| mTOR signaling pathway                   | 19 (11.45) | 12 | 8 (11.43) | 6 | 9 (10.23) | 8 | 6 (9.68) | 4 | 7 (9.21) | 5 | 13 (14.94) | 8 |
| ECM-receptor interaction                 | 17 | 10 | 7  | 5 | 7  | 6 | 6 | 4 | 7  | 5 | 13 | 8 |
| Cell adhesion molecules                  | 3  | 2  | 1  | 1 | 2  | 2 | 0 | 0 | 2  | 0 | 2  | 1 |
| Cellular Processes                       | 152 | 85 | 68 | 41 | 70 | 42 | 54 | 4 | 64 | 44 | 83 | 56 |
| Cellular community                       | 54 (35.53) | 27 | 27 (39.71) | 13 | 29 (41.43) | 15 | 26 (48.15) | 16 | 27 (42.19) | 16 | 36 (43.37) | 21 |
| Focal adhesion                           | 35 | 21 | 13 | 10 | 17 | 11 | 14 | 11 | 15 | 11 | 23 | 17 |
| Tight junction                           | 19 | 10 | 18 | 6 | 17 | 6 | 15 | 8 | 16 | 7 | 10 | 6 |
| Gap junction                             | 8  | 2  | 1  | 1 | 1  | 1 | 2 | 2 | 2  | 2 | 7  | 2 |
| Adherens junction                        | 5  | 3  | 4  | 3 | 5  | 3 | 3 | 3 | 5  | 3 | 3  | 3 |
| Transport and Catabolism                 | 42 (27.63) | 24 | 17 (25.00) | 11 | 20 (28.57) | 14 | 8 (17.81) | 8 | 16 (25.00) | 11 | 18 (21.69) | 11 |
| Cell Growth and Death                    | 36 (23.68) | 22 | 16 (23.53) | 11 | 12 (14.87) | 8 | 13 (24.07) | 12 | 14 (21.88) | 13 | 19 (22.89) | 16 |
| Apoptosis                                | 19 | 13 | 13 | 7 | 5  | 4 | 6 | 6 | 7  | 7 | 11 | 9 |
| p53 signaling pathway                    | 7  | 5  | 6  | 4 | 3  | 1 | 1 | 1 | 3  | 3 | 4  | 4 |
| Cell Motility                            | 20 (13.16) | 12 | 8 (11.76) | 6 | 9 (12.86) | 5 | 7 (12.96) | 6 | 7 (10.94) | 4 | 10 (12.05) | 8 |
| Regulation of actin cytoskeleton         | 20 | 12 | 8  | 6 | 9  | 5 | 7 | 6 | 7  | 4 | 10 | 8 |
| Organismal Systems                       | 274 | 154 | 108 | 66 | 124 | 84 | 109 | 81 | 122 | 82 | 129 | 89 |
| Endocrine System                         | 105 (38.32) | 53 | 44 (40.74) | 25 | 49 (39.52) | 32 | 36 (33.03) | 24 | 53 (43.44) | 3 | 56 (43.41) | 33 |
| Glucagon signaling pathway               | 36 | 12 | 19 | 8 | 14 | 9 | 12 | 7 | 22 | 11 | 23 | 10 |
| Insulin signaling pathway                | 32 | 12 | 14 | 7 | 11 | 6 | 7 | 4 | 12 | 6 | 22 | 10 |
| Thyroid hormone signaling pathway        | 11 | 7  | 4  | 4 | 6  | 5 | 3 | 3 | 4  | 4 | 6  | 6 |
| Thyroid hormone synthesis                | 6  | 4  | 3  | 2 | 1  | 1 | 3 | 3 | 2  | 2 | 1  | 1 |
| Circulatory System                       | 49 (17.88) | 30 | 18 (16.67) | 12 | 22 (17.74) | 15 | 19 (17.43) | 16 | 16 (13.11) | 13 | 15 (11.63) | 12 |
| Immune System                            | 44 (16.06) | 28 | 16 (17.81) | 10 | 21 (16.94) | 14 | 22 (20.18) | 17 | 24 (19.67) | 16 | 23 (17.83) | 17 |
|                          | Digestive System | Nervous System | Aging | Total |
|--------------------------|------------------|----------------|-------|-------|
|                          | 32 (11.68)       | 27 (9.85)      | 17 (6.20) | 1043 | 632 | 448 | 301 | 491 | 342 | 377 | 285 | 470 | 332 | 515 | 363 |
|             Protein digestion and absorption | 12 5 6 | 5 12 (9.68) | 5 | 10 10 (9.26) | 5 | 285 | 470 | 332 | 515 | 363 |
|                          | 9 9 (7.26)       | 9 11 (10.09)   | 7 | 3 2 (7.5) | 3 | 342 | 377 | 285 | 470 | 332 | 515 | 363 |
|                          | 7 18 (16.54)     | 10 10 (8.20)   | 3 | 6 (4.92) | 2 | 3 2 (7.5) | 3 | 342 | 377 | 285 | 470 | 332 | 515 | 363 |
|                          | 11 13 (10.66)    | 9 9 (6.98)     | 2 | 6 (4.65) | 5 | 9 9 (6.98) | 8 | 10 10 (8.20) | 9 | 9 (6.98) | 8 | 3 2 (7.5) | 3 | 342 | 377 | 285 | 470 | 332 | 515 | 363 |
|                          | 9 20 (15.50)     | 10 10 (8.20)   | 9 | 9 (6.98) | 8 | 10 10 (8.20) | 9 | 9 (6.98) | 8 | 3 2 (7.5) | 3 | 342 | 377 | 285 | 470 | 332 | 515 | 363 |
|                          | 14               | 8              | 5 | 5 (4.65) | 2 | 9 (6.98) | 8 | 3 2 (7.5) | 3 | 342 | 377 | 285 | 470 | 332 | 515 | 363 |
transcripts/18 KOs belonging to oxidative phosphorylation. Shear force exhibited an increased number of transcripts belonging to lipid metabolism (16 transcripts, 14 KOs) (Table 4).

Our KEGG pathway mapping results have linked many of the genes harboring SNPs with allelic imbalances to potential regulation of growth and metabolic pathways, which may support pathway-based GWAS analyses in rainbow trout, similar to what has been recently applied to detect genetic pathways explaining live weight and muscle growth variation in cattle genotypes [52].

Methods
Fish population, sampling and sequencing
Phenotypic data and muscle samples were collected from ~500 fish representing 98 families (5 fish/family) from the growth-selected line at NCCCWA (year class 2010) as previously described [6, 11, 26]. Families were produced and reared until ~13 months post-hatch as described in reference [26]. Briefly, full-sib families were produced from single-sire × single-dam matings. Eggs were reared in spring water, and water temperatures were manipulated between approximately 7 and 13 °C to synchronize hatch times. Each family was stocked separately in 200-L tanks at a density of approximately 600 alevins/tank. Fish were randomly culled every month to maintain stocking densities <50 kg/m3. At about 5-months old, fish were anesthetized using 100 mg/L of tricaine methanesulfonate (Tricaine-S, Western Chemical, Ferndale, WA) and uniquely tagged by inserting a passive integrated transponder (Avid Identification Systems Inc., Norco, CA) into the dorsal musculature, parallel to the long axis of the body, from the dorsal musculature for texture analysis [54]. The remaining muscle from the fillets was pulverized with liquid nitrogen in a Waring Blender (Waring, New Hartford, CT) and kept at −25 °C for chemical composition analyses. Proximate composition of muscle was determined using AOAC [55] approved methods. Crude fat was analyzed using the Soxhlet solvent extractor with petroleum ether. Texture of fillet sections was determined using a five-blade, Allo-Kramer shear cell attached to a Texture Analyzer (Model TA-HDi®; Texture Technologies Corp., Scarsdale, NY), equipped with a 50-kg load cell and at a crosshead speed of 127 mm/min. Force-deformation graphs were recorded and analyzed using the Texture Expert Exceed software (version 2.60; Stable Micro Systems Ltd., Surrey, U.K.). Peak shear force (g/g sample) was recorded.

Fresh fillet surface color was measured with a Chromameter (Minolta, Model CR-300; Minolta Camera Co., Osaka, Japan) calibrated using a standard white plate No. 21333180 (CIE Y 93.1; x 0.3161; y 0.3326). L* (lightness), a* (redness), and b* (yellowness) values were recorded at three locations above the lateral line along the long axis of the right fillet, and these values were used to calculate a fillet whiteness index according to the following equation:

\[
\text{Whiteness} = 100 - \left(\frac{100 - L^*}{a^* + b^*}\right)^{1/2}
\]

For RNA-Seq analyses, out of 98 families measured for phenotypic data, eight families (5 fish each) showing opposite phenotypes for each of the 5 traits were analyzed (4 high ranked families versus 4 low ranked families on average for each trait). Since some fish families were common between the traits, the total number of selected families for RNA-Seq was 22 families. Total RNA was isolated from each fish muscle sample using TRIzol™ (Invitrogen, Carlsbad, CA). Equal masses of total RNA from 5 samples of each family were pooled and used for RNA-Seq sequencing. cDNA libraries were prepared and sequenced on an Illumina HiSeq (single-end, 100 bp read length) using multiplexing standard protocols as previously described [56]. Briefly, mRNA was selected
from one microgram of high quality total RNA. First-strand synthesis was synthesized with a random hexamer and SuperScript II (Life Technologies). Double stranded DNA was blunt-ended, 3’-end A-tailed and ligated to indexed adaptors. The adapter-ligated double-stranded cDNA was amplified by PCR for 10 cycles with the Kapa HiFi polymerase (Kapa Biosystems, Woburn, MA) to reduce the likelihood of multiple identical reads due to preferential amplification. The final libraries were quantitated Qubit (Life Technologies, Grand Island, NY) and the average size was determined on an Agilent bioanalyzer DNA7500 DNA chip (Agilent Technologies, Wilmington, DE), diluted to 10 nM and the indexed libraries were pooled in equimolar concentration before sequencing.

**SNP detections using SAMtools/Popoolation2**

For each trait (WBW, muscle yield, muscle fat content, shear force, and whiteness), sequence reads from each family were aligned to the rainbow trout genome using STAR [29]. After read alignment, the SAMtools view/sort and mpileup functions were used within the Popoolation2 package (version 1.201) to determine the genotype for each variant and calculate allele frequencies [57, 58]. Initial SNPs were considered at minimum reads >10 and minor allele count >4 and MAF > 0.05. Putative SNPs associated with each trait were determined by calculating SNP allelic imbalance scores as previously described [11, 59]. A SNP allelic imbalance score was determined by assessing the ratio of [frequency of allele A/frequency of allele B in high-end families]/[frequency of allele A/frequency of allele B in the corresponding low-end families]. The allelic imbalances score ranges from zero to infinity. SNPs with allelic imbalance were called if the ratio is more than or equal 2.0 (as an amplification) or less than or equal 0.5 (as loss of heterozygosity). The phase of the alleles could not be determined for families surveyed since the parental genotypes were not known for most of the fish. Allele counts in the divergent families were extracted from the VCF files. Chi-square test of two-by-two Tables [60] was performed with p-value <0.05 to determine if SNPs that are showing allelic imbalances are statistically significant.

**SNP detection using GATK tools**

For the GATK pipeline [61], reads from each sample were aligned to the rainbow trout genome using STAR [29] as recommended by the GATK practice. Picard tools were used to sort the SAM files and to mark duplicates, a step used by GATK to reduce a false positive due to error in duplicate that could be falsely detected as a SNP. The following steps were performed according to GATK pipeline for RNA-Seq (Split and trim to reassign mapping quality, Indel realignment, local realignment around Indel in order to clean up any mapping artifacts and Base Quality Score Recalibration). After data preparation, variants were called using HaplotypeCaller followed by hard-filtering using the following parameters: Qual By Depth (QD) 2.0, FisherStrand (FS) 60.0: RMS Mapping Quality (MQ) 40.0, MAF > 0.05. Since GATK was not optimized to calculate allelic imbalances in RNA-Seq data, putative SNPs identified in each family were analyzed using an in-house Perl script to determine the allelic imbalances applying the criteria that we used in the SAMtools/Popoolation2 method.

**SNP validation**

Flanking sequences (up to 250 bp on each side) of putative SNPs were extracted from the reference genome [28]. Some SNPs were removed from SNP assay design because either a sequence gap was located less than 60 bp from the SNP site or a non-target SNP was located less than 30 bp away from the target SNP. A total of 92 SNP assays were developed and evaluated with 282 DNA or cDNA samples. These included 85 DNA samples derived from 19 full-sib families used for RNA-Seq and their parents (38 DNA samples), DNA samples of 2 full-sib mapping families (2 parents and 19 offspring per family), 64 DNA samples from two commercial populations (Troutlodge Inc. and Clear Springs Foods Inc.) and 35 cDNA samples derived from the RNA samples used for RNA-Seq high versus low muscle yield. The SNP genotyping was performed following the instructions of the Fluidigm genotyping user guide. Briefly, DNA and cDNA samples were pre-amplified, diluted and used for genotyping with 96.96 Dynamic Array IFCs (Integrated Fluidic Circuits). The arrays were read using EP1 system, and genotypes were called automatically using Fluidigm SNP genotyping analysis software 4.1 with a confidence threshold of 85. The genotype clusters were examined for each assay and any wrong calls or no calls were corrected manually. The program Pedcheck [62] was used to identify genotypes inconsistent with Mendelian inheritance between parents and offspring. Chi-square goodness of fit tests were performed to identify SNPs with significant segregation distortion (P < 0.01) in the two mapping families. Those SNPs were reported as assay-failed SNPs.

For the Sanger sequencing validation of the SNPs showing potential mono-allelic gene expression, flanking sequences (up to 250 bp on each side) of each SNP were PCR amplified from DNAs and cDNA from the same 35 fish samples that were used for RNA-Seq high versus low muscle analyses. PCR amplicons were Sanger sequenced and manually inspected for consistency between DNA and cDNA genotypes or mono-allele specific gene expression as explained in the results section.
Functional annotation of SNPs
SNP annotation by functional class (genic/intergenic etc.) for different SNP sets and their genome distributions were conducted using in-house Perl scripts. The gff file of the rainbow trout genome reference [28] was used to determine if a SNP is located within an mRNA start and end positions (genic), within a CDS, 5’UTR or 3’UTR. SNPs not within start and end positions of mRNA were considered intergenic. Upstream/downstream intergenic SNPs were determined if located within 5 kb of an mRNA. SNPs within IncRNAs were determined using gff file of our previously reported IncRNA reference [63]. Functional annotation of the SNP-harboring genes was performed using the Blast2GO suite [30] and KEGG pathway mapping.

Additional file

Additional file 1: Putative SNPs and SNPs with allelic imbalances in association with total body weight, muscle yield, muscle fat content, shear force, and whiteness. SNP chromosome position, alleles, functional classification, associated gene ID and FASTA sequences are provided. (CSV 15332 kb)

Abbreviations
CDS: Coding DNA sequences; FS: Fisher Strand; GATK: Genome Analysis Toolkit v3.3.0; GO: Gene Ontology; KOs: KEGG Orthologies; MAF: Minor allele frequency; MQ: Mapping Quality; PSE: Pale, soft, and exudative; QD: Qual By Depth; WBW: Whole body weight

Acknowledgments
Not applicable.

Funding
This study was supported by a competitive grant No. 2014–67,015-21,602 from the United States Department of Agriculture, National Institute of Food and Agriculture (MS).

Availability of data and materials
All SNP data are provided in additional file 1 and were deposited in the NCBI dbSNP database (accession number from ss#2711191806 to 2,711,287,038 in addition to ss#2137497773). While the manuscript was under review for publication, a newer assembly of the rainbow trout genome was deposited at NCBI (GenBank: MSJN00000000.1, Dr. Paliti personal communication). SNP flanking sequences were mapped to the new reference and positions of the uniquely mapped SNP sequences were annotated in Additional file 1.

Authors’ contributions
Conceived and designed the experiments: MS, TL, BK. Performed the experiments: RA, MS, TL, BK. Analyzed the data: RA, AA, SL, YP, BK, MS. Wrote the paper: MS, RA, AA. All authors reviewed and approved the publication.

Ethics
Institutional Animal Care and Use Committee of the United States Department of Agriculture, National Center for Cool and Cold Water Aquaculture (Leetown, WV) specifically reviewed and approved all husbandry practices used in this study (IACUC approval #056).

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
1Computational Science Program, Middle Tennessee State University, Murfreesboro, TN 37132, USA. 2Department of Biology and Molecular Biosciences Program, Middle Tennessee State University, Murfreesboro, TN 37132, USA. 3National Center for Cool and Cold Water Aquaculture, ARS-USDA, Kearneysville, WV 25430, USA. 4Division of Animal and Nutritional Sciences, West Virginia University, Morgantown, WV 26506, USA.

Received: 26 December 2016 Accepted: 1 August 2017

Published online: 07 August 2017

References
1. Salem M, Kenney PB, Rexroad CE, Yao J. Molecular characterization of muscle atrophy and proteolysis associated with spawning in rainbow trout. Comp Biochem Physiol Part D Genomics Proteomics. 2006;1(2):227–37.
2. Gjedrem T. Selection and breeding programs in aquaculture. New York: Springer; 2008.
3. Rexroad CE, Paliti Y, Gahr SA, Vallego RL. A second generation genetic map for rainbow trout (Oncorhynchus Mykiss). BMC Genet. 2008;9:74–4.
4. Wang P, Sun L, Bao L, Zhang J, Jiang Y, Yao J, Song L, Feng J, Liu S, Liu Z. Bulk segregant RNA-seq reveals expression and positional candidate genes and allele-specific expression for disease resistance against enteric septicemia of catfish. BMC Genomics. 2013;14:929.
5. Wang S, Sha Z, Sonstegard TS, Liu H, Xu P, Somridthive B, Peatman E, Kucuktas H, Liu Z. Quality assessment parameters for EST-derived SNPs from catfish. BMC Genomics. 2008;9:945.
6. Gonzalez-Pena D, Gao G, Baranski, M., Moen, T., Cleveland B, Kenney P, Vallejo R, Paliti Y, Leeds T. Genome-wide association study for identifying loci that affect fillet yield, carcass, and body weight traits in rainbow trout (Oncorhynchus Mykiss). . Front Genet 2016, 7.
7. Paliti Y, Gao Y, Liu S, Kent MP, Lien S, Miller MR, Rexroad CE, Moen T. The development and characterization of a 5’S single nucleotide polymorphism array for rainbow trout. Mol Ecol Resour. 2015;15(3):662–72.
8. Tsai HY, Hamilton A, Tinche AE, Guy DR, Gharbi K, Stear MJ, Matika O, Bishop SC, Houston RD. Genome wide association and genomic prediction for growth traits in juvenile farmed Atlantic salmon using a high density SNP array. BMC Genomics. 2015;16:969.
9. Caffton VE, Ireland JS, Useche F, Fahlam M. Functional single nucleotide polymorphism-based association studies. Hum Genomics. 2006;2(6):391–402.
10. Brookes AJ. Single Nucleotide Polymorphism (SNP). In: ENCYCLOPEDIA OF LIFE SCIENCES (ed). Edited by John Wiley & Sons Lwen, http://onlinelibrary.wiley.com/doi/10.1002/9780470015902.a0005006.pub2/full, accessed April 4, 2012; 2007.
11. Salem M, Vallego RL, Leeds TD, Paliti Y, Liu S, Sabbagh A, Rexroad CE, Yao J. RNA-Seq identifies SNP markers for growth traits in rainbow trout. PLoS One. 2012;7(5):e36264.
12. Villanueva B, Dekkers JC, Woudaamn JA, Sartor P. Maximizing genetic gain over multiple generations with quantitative trait locus selection and control of inbreeding. J Anim Sci. 2004;82(S):1305–14.
13. Dekkers JC. Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. J Anim Sci 2004, 82 E-Suppl:E33-E328.
14. Pang Y, Wang J, Zhang C, Lei C, Lan X, Yue W, Gu C, Chen D, Chen H. The polymorphisms of bovine VEGF gene and their associations with growth traits in Chinese cattle. Mol Biol Rep. 2010.
15. Tsai HY, Hamilton A, Guy DR, Tinche AE, Bishop SC, Houston RD. The genetic architecture of growth and fillet traits in farmed Atlantic salmon (Salmo Salar). BMC Genet. 2015;16:61.
16. Harvey DJ. Aquaculture Outlook. Electronic Outlook Report from the Economic Research Service www.ers.usda.gov; In: 2006.
17. Davidson WS. Adaptation genomics: next generation sequencing reveals a shared haplotype for rapid early development in geographically and genetically distant populations of rainbow trout. Mol Ecol. 2012;21(2):219–22.
18. Speare D, Arnsenau G, Buote M. Evaluation of rainbow trout as a model for use in studies on pathogenesis of the branchial microsporidian Loma salmonae. Contemp Top Lab Anim Sci. 1998;37(2):55–8.
19. Salem M, Kenney PB, R rexord CE 3rd, Yao J. Microarray gene expression analysis in atrophying rainbow trout muscle: a unique nonmammalian muscle degradation model. Physiol Genomics. 2006;28(1):133–45.

20. Patel M, Rogers JT, Pane EF, Wood CM. Renal responses to acute lead waterborne exposure in the freshwater rainbow trout (Oncorhynchus Mykiss). Aquat Toxicol. 2006;80(4):362–71.

21. Thorgaard GH, Bailey GS, Williams D, Buhler DR, Kaatari SL, Ristow SS, Hansen JD, Winton JR, Bartholomew JI, Nagler JJ, et al. Status and opportunities for genomics research with rainbow trout. Comparative biochemistry and physiology Part B, Biochemistry & molecular biology. 2002;133(4):609–46.

22. Crulli ET, Singh A, Shanna KV, Ge D, Smith JP, Maia JM, Heinzén EL, Goedert JJ, Goldstein DB. Center for HIVAVI: screening the human exome: a comparison of whole genome and whole transcriptome sequencing. Genome Biol. 2010;11(5):R57.

23. Heap GA, Yang JH, Downes K, Healy BC, Hunt KA, Bockett N, Franke L, Dubois PC, Mein CA, Dobson RJ, et al. Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. Hum Mol Genet. 2010;19(1):122–34.

24. Vidal RO, do Nascimento LC, Monteiro JM, Pereira GA, Carazzone MF. Identification of SNPs in RNA-seq data of two cultivars of Glycine max (soybean) differing in drought resistance. Genet Mol Biol. 2012, 35(3):331–344.

25. Yang SS, Tu ZL, Cheung F, Xu WW, Lamb JF, Jhung HJ, Vance CP, Gronwald JW. Using RNA-Seq for gene identification, polymorphism detection and transcript profiling in two alfalfa genotypes with divergent cell wall composition in stems. BMC Genomics. 2011;12:2199.

26. Leeds TD, Vallejo RL, Weber GM, Pena DG, Silverstein JS. Response to five generations of selection for growth performance traits in rainbow trout (Oncorhynchus Mykiss). Aquaculture. 2016;465:341–51.

27. Leeds TD, Kenney PB, Manor M. Genetic parameter estimates for feed intake, body composition, and fillet quality traits in a rainbow trout population selected for improved growth. In: International Symposium on Genetics in Aquaculture 2012, Auburn, AL, 2012: 259.

28. Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noel B, Bento P, Da Silva C, Fabadie K, Alberti A, et al. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat Commun. 2014;5:3657.

29. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.

30. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Grosskopf S, Roy J, D’Agostino D, Fennell T, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303.

31. Raineri E, Ferrari L, Esteve-Codina A, Nevado B, Heath S, Pérez-Enciso M. SNP polymorphisms for rainbow trout generated by restriction-site associated sequencing of a reduced representation library. BMC Genomics. 2009;10:559.

32. Fan H, Wu Y, Zhou X, Jia J, Wang Z, Song Y, Liu F, Chen Y, Zhang L, Gao X, et al. Pathway-based genome-wide association studies for two meat production traits in Simmental cattle. Sci Rep. 2015;5:18389.

33. Sanchez CC, Smith TP, Wiesmann RT, Vallejo RL, Salem M, Yao J, Rexord CE 3rd. Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. BMC Genomics. 2009;10:559.

34. Piskol R, Ramaswami D, Reinus B, Sandberg R. Single-cell RNA-seq reveals dynamic, random mononuclear gene expression in mammalian cells. Science. 2014;343(6167):193–6.

35. Piskol R, Ramaswami G, Li JB. Reliable identification of genomic variants from RNA-seq data. Am J Hum Genet. 2013;93(4):641–51.

36. Danzmann RG, Kocmarek AL, Norman JD, Rexord CE, Palti Y. Transcriptome profiling in fast versus slow-growing rainbow trout across seasonal gradients. BMC Genomics. 2016;17.

37. Wringe BF, Devlin RH, Ferguson WM, Moghadam HK, Sahkhan D, Danzmann RG. Growth-related quantitative trait loci in domestic and wild rainbow trout (Oncorhynchus Mykiss). BMC Genet. 2010;11.

38. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SN, Nueda MJ, Robles M, Talon M, Dopazo J, Conesa A. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 2008;36(10):3430–35.

39. Salem M, Kenney PB, R rexord CE 3rd, Yao J. Proteomic signature of muscle atrophy in rainbow trout. J Proteome. 2010;73(4):778–89.

40. Salem M, Kenney PB, Rexord CE III, Yao J. Development of a 37 k high-density oligonucleotide microarray: a new tool for functional genome research in rainbow trout. J Fish Biol. 2008;72(2):2187–206.

41. Deng Q, Ramskold D, Reinius B, Sandberg R. Single-cell RNA-seq reveals allele-specific expression in the interspecific F1 hybrid catfish. Mol Gen Genomics. 2016;393(1):1083–93.

42. Piskol R, Ramaswami D, Reinus B, Sandberg R. Single-cell RNA-seq reveals dynamic, random mononuclear gene expression in mammalian cells. Science. 2014;343(6167):193–6.

43. Salem M, Kenney PB, R rexord CE 3rd. Microarray gene expression one allele at a time. Trends Genet. 2014;30(6):237–44.