Elucidating the molecular nature of electric organs using genomic, transcriptomic, and proteomic approaches.

By:

Lindsay L. Traeger

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy (Genetics)

at the

UNIVERSITY OF WISCONSIN-MADISON

2015

Date of final oral examination: October 22, 2015 at 2 pm

The dissertation is approved by the following members of the Final Oral Committee:

Dr. Michael R. Sussman, Professor of Biochemistry

Dr. Audrey P. Gasch, Associate Professor of Genetics

Dr. Garret Suen, Assistant Professor of Bacteriology

Dr. Francisco Pelegri, Professor of Genetics and Medical Genetics

Dr. Allen Laughon, Professor of Genetics and Medical Genetics

Table of Contents

Abstract

Electric fishes have played important roles historically in science, helping to elucidate the very nature of electricity, and in determining the first structures and functions of ion transporters. Despite their historical importance, no study has comprehensively determined the molecular nature of electric organs (EOs), nor have any studies compared EOs in distinct lineages of electric fishes on a large scale. Hundreds of species of electric fish have been described across diverse taxa, the vast majority of which are capable of producing only weakly electric organ discharges (EODs) for the purposes of navigation and communication. One such taxa, the Gymnotiformes, are native to South American fresh waters, and show striking diversity in their weak EO signals. For example, the strong-voltage electric eel (*Electrophorus electricus*) is unique among the Gymnotiformes as not only does it have the ability to generate weak EODs, it also has the incredible ability to generate high-voltage discharges (in excess of 600 volts) for the purposes of predation and defense. In this dissertation, I characterized the genome of *E. electricus* using high-throughput DNA sequencing technologies. I also sequenced mRNA and miRNA transcripts obtained from eight tissues, including the three EOs found in this species and skeletal muscle. Next, I used comparative transcriptomics to study the sequence and abundance of transcripts found in EO and muscle in two additional Gymnotiformes species, and in two lineages of fish that have independently evolved EOs. We found similar patterns of gene expression in the EOs of these fishes despite their independent origins, notably of transcription factors and signaling pathways that may represent a common genetic toolkit important for the independent evolution of EOs. Finally, I used a quantitative proteomic and phosphoproteomic approach to characterize

each of the three distant EOs in *E. electricus*, and found differences among the EOs that I have hypothesized may reflect their distinct functions. The main EO (strong voltage) is used intermittently and is most abundant in proteins vital to EOD: acetylcholine receptor and acetylcholinesterase, voltage-gated sodium channel, and Na^+/K^+-ATPase. In contrast, the Sachs' EO (weak-voltage) is used continuously, and is least abundant in these proteins. Together, the differences in protein abundance and phosphorylation of these key proteins in the electrogenic machinery may represent the molecular underpinning between the tissues' different energetic needs– generating low voltages continuously versus generating high voltages intermittently. I also identified differences in signaling pathways and transcription factors between the distinct EOs that may provide clues to differences in their development and/or function. In summary, this dissertation represents the first major effort at characterizing a novel tissue type, the EO, both within a species and across diverse species.

Acknowledgements

Thank you to my mentor, Mike, for being a constant and simultaneous example of novel idea generation and careful science, for providing an endless stream of jokes, for creating an environment in which the people in your lab enjoy each others company both in and out of the lab, and for being a willing sponsor and participant of Friday Club. I have felt incredibly lucky and incredibly supported by you and the environment that you have created for the people in your lab. Thank you for allowing me to be a part of it.

Thank you to my committee members, Audrey Gasch, Al Laughon, Francisco Pelegri, and Garret Suen, for your support and many thoughtful discussions over my years as a graduate student, both in my committee meetings and out.

Thank you to all of our collaborators- the Eelistas- for collectively being an endless source of knowledge on 'all things electric.'

Thank you to the members of the Sussman Lab and Mass Spectrometry Facility. Collectively, you are and endless wealth of knowledge. In addition to colleagues you have become my friends. A special thanks to Jeremy Volkening, for teaching me the ropes in bioinformatics, and to Kelly Stecker, Heather Burch, and Melanie Ivancic, for our many discussions in all things science and all things otherwise.

Thank you to my family, for being endlessly supportive of me throughout my life, for giving me unconditional love, and for absolutely always being in my corner. Thank you to my mother, for being a demonstration of resilience and strength, for teaching me to stop and smell the roses. Thank you for my father, for your endless creativity, your logic. Thank you to my brother, for being a demonstration of how to work and play hard, and for fielding my computer-related questions.

Finally, thank you to my husband. You are always have my back. Thank you for all of your support in the big things and the small things, for your endless patience, your thoughtfulness, and for always trying to make me laugh, even when I really don't want to. Also, thank you for reading this entire dissertation.

List of Figures

List of Tables

Chapter 1: A historical perspective of electric fishes in culture and science.

Or arm in waves, electric in his ire,

The dread Gymnotus with ethereal fire.—

Onward his course with waving tail he helms,

And mimic lightnings scare the watery realms. ...

-Erasmus Darwin (1731-1802), in Canto 1 of *Economy of Vegetation*

Introduction to electric fish in history

Electric fishes have captured the imagination of humans for millennia. References to an electric fish date back to 5000 years ago, in ancient Egypt where images of the strong-voltage electric catfish *Malapterurus electricus* were used as a hieroglyph to refer to a "man who has saved many in the sea [1]." This was because any fisherman who caught an electric catfish together in their net with other fish would be shocked through the wet net or pole, causing them pain and ultimately making them drop the net, liberating their catch.

The curiosity sparked in mankind by electric fishes ultimately shaped the world we live in. Electric fishes have played a critical role in physics, helping to elucidate the nature of electricity. For example, the invention of the Leyden Jar in the 1700s which was used to store and transport electricity generated by electrostatic generators and the numbing sensation experienced from the Leyden jar caused scientists to draw comparisons with the numbing sensation experienced from handling *Torpedo* rays and electric eels [1]. By 1800, Alessandro Volta had used the electric organs in *Torpedo* and the electric eel as inspiration towards the design of an "artificial electric organ," the very first electric battery (the Voltaic pile) [1, 2]. This consisted of alternating zinc and silver disks followed by 'spongy matter' such as leather, all soaked in saline solution.

Electric fishes have also played a vital role in biochemistry and neuroscience, helping scientists first understand how ions could move across cell membranes [3]. Electric organ from the strong-voltage *Torpedo* ray was critical for the discovery of the neurotransmitter acetylcholine and cholinesterase [4], and provided such an abundant source of the acetylcholine receptor (a sodium channel) that the electric organ of *Torpedo* was the source for the first ion channel ever cloned and sequenced [3, 5, 6]. Similarly, electric organs of the strong-voltage electric eel (*Electrophorus electricus*) are incredibly rich in voltage-gated sodium channels, and were used to first clone and sequence this protein [5, 7]. Electric organs of the *E. electricus* are also the richest known source for the Na^+/K^+ ATPase, and the first biochemical analyses of this enzyme were done using the electric eel [8, 9].

Despite the important historical significance of electric fishes, relatively little is known about their molecular nature, and few studies have offered an analysis of their molecular nature or evolution.

What is known about electric organs of other species?

The ability to perceive weak electric fields (electroreception) is an ancestral vertebrate trait and because water conducts electricity while air is an insulator, the ability to emit electric discharges, as well as the ability to detect small changes in the electric fields in their environment, are capabilities only found in organisms that dwell (at least partially) in water [10]. Electroreceptors are found in most non-teleost fishes, in and four orders of teleost fishes (all *Siluriformes*, *Gymnotiformes*, *Mormyriforms*, and one subfamily of the *Osteoglosiformes*), in some amphibians, and on the bill of the *Platypus.* Unlike

electroreception, electrogeneration—the ability to produce coordinated electric fields from electric organs— is not considered ancestral and is thought to have evolved independently at least six times in fishes, possibly to complement existing electroreceptive ability (Figure 4.1, A) [11-13]. Among cartilaginous fishes, electric organs are thought to have evolved twice-- in the Torpedinoids (electric rays, ~38 species) and in the Rajoids (electric skates, >200 species). In teleost fishes, electric organs are thought to have evolved at least four times, in Mormyroids (elephant noses, ~200 species), the Siluriformes (catfishes, several of the 34 families), the stargazers (two of the 148 families of Perciformes) and all Gymnitoformes (knifefishes, ~200 species).

In this chapter I provide a brief overview about the electric organs of all electric fish lineages, but will focus mainly on the electric organs of Gymnotiformes, and in particular, the electric eel.

Electric Organs in the Cartilaginous Fishes

Torpedinoids (electric rays): The electric rays have paired kidney-shaped electric organs located in the pectoral muscles that develop from branchiomeric musculature [14]. Using these electric organs, electric rays can generate ~50 V and 50 A of current [15]. The electric organ discharges from the electric rays are used in nocturnal hunting and are strong enough to stun and occasionally cause tetanus so powerful as to snap the vertebrae of its prey [16]. Despite their historical importance in neuroscience (see introductory section for a brief overview), few studies have performed an in-depth molecular analysis of their electric organs. A small number experiments aimed to identify proteins in the neuromuscular junction of the electric organ of the *Torpedo* ray [15, 17]. Maximally, these

studies identified ~400 proteins, and did not do a comparison to *Torpedo* muscle directly, but rather, to mouse skeletal muscle [15].

Rajoids (electric skates): Electric skates, like the electric rays, possess a paired electric organ; however, unlike that of the electric ray, the electric organs develop from hypaxial skeletal muscle fibers, and are spindle-shaped and confined to the tail along each side of the vertebral column from within the tail [18]. These electric organs emit only weak electric organ discharges that are used in communication only, and surprisingly, do not have a known role in navigation or defense.

Electric Organs in the Teleost Fishes

Uranoscopidae (the Stargazers): The stargazers belonging to the genera *Astroscopus* and *Uranoscopus* are the only known marine electric teleosts [19, 20]. There are a small number of species belonging to two genera that are electric [21]. The electric organs in stargazers are derived from ocular muscles [22] and can generate upwards of ~50V. Although the function of their electric organs remains unclear because their electric organ discharges are too weak to stun prey, they are likely used in predation and defense because agitation of the fish by poking with a stick elicits a predictable electric organ discharge [19, 21], but because the species lack any known electrosensory system (i.e., no electroreceptors), their electric organs are likely not used in communication or navigation [19, 21].

Siluriformes (catfishes): Catfish belonging to the genus *Malapterurus* are the only strong-voltage fish that use electric organ discharges for communication as well as predation and

defense (no other fish identified to date use their strong voltage organs for communication purposes) [20]. Electric organs in *Malapterurus* are paired and develop from trunk muscles, covering most of the body in a sheath that runs under just under the skin from just behind the head to nearly as far as the anal fin [23] Dissimilarly, electric catfish belonging to the genera *Claria* and *Syndontis* have weakly electric organs that are used in communication [20]. Interestingly, *Syndontis* has electric organs that develop from sonic muscle associated with the swim bladder [20]. Other catfish belonging to the genera *Auchenoglanis*, *Ompok*, and *Ictalurus* have been reported to have electric organ discharges from unknown origin, and the vastly different tissue origins for electric organs in catfish has lead to hypotheses that within catfish multiple independent evolutionary origins of electric organs have occurred, but these claims have yet to be confirmed.

Mormyroids (African electric fish): The Mormyrids have a weak, myogenically-derived electric organ used in navigation and communication that is derived from axial mesoderm; in adults, the electric organ is small and restricted to the caudal peduncle of the tail [20]. Group hunting has been demonstrated in mormyroid fish in which occasionally electric organ discharges were synchronized, which suggests it may function as a pack cohesion signal in complex communication behaviors [24].

Gymnotiformes (South-American electric fish): Nearly 200 different species of Gymnotiformes have been identified to date, all native to the South American neotropics [20]. Gymnotiformes are quite diverse anatomically, from the electric organ discharges, and from the electric organ structures themselves. Most Gymnotiformes have electric

organs that are derived from skeletal muscle precursors; adults in species belonging to *Apternotu*s have neurogenic electric organs. *Apternotus* begins life with a myogenic ELECTRIC ORGAN, but in 13-day-old larvae, the neurogenic organ begins to appear which are modified axons of electromotor neurons; by 41 days, the influence of the myogenic organ on electric organ discharge is ablated [25]. *E. electricus* is unique among the Gymnotiformes as it has three distinct electric organs (Figure 3.1), and has the ability to produce both strong and weak-voltage discharges -- all other species have weakly electric myogenic or neurogenic organs.

How do electric organs function?

Electric organs are comprised of individual cells called electrocytes, which are specialized in generating ion flow. In the *E. electricus*, individual electrocytes are made up of two 'faces:' the innervated face contains acetylcholine receptors (ligand-gated Na^+ channel) and voltage-gated Na^+ channels, and thus, is both chemically and electrically excitable; the opposite membrane face is not innervated and is also not excitable, and instead of Na^+ channels, contains Na^+/K^+ ATPase (Figure 1.1, A) [5, 26]. This is in contrast to the marine electric fish species such as *Torpedo*, which are not electrically excitable, and instead rely entirely on the acetylcholine receptor to trigger membrane depolarization [26] To activate electrocytes in *E. electricus*, acetylcholine is released from the innervating neuron and binds to the acetylcholine receptor, allowing a small amount of Na^+ to flow into the cell, which causes a slight membrane depolarization. This initial membrane depolarization in turn causes the voltage-gated Na^+ channels to open, causing a massive influx of Na^+ across the innervated membrane face [26]. Because the opposite membrane

face contains no Na^+ channels, a large trans-cellular potential difference (voltage) is generated in each individual electrocyte (~150 mV per electrocyte), and because electrocytes are arranged massively in series in an electric organ (like batteries in a flashlight), the individual voltages sum up (Figure 1.1,B). Further, a high voltage (e.g., 600 volts in a mature eel) is generated because electrocytes are also arranged massively in parallel, spanning the width of the tail of an electric eel. The presence of Na+/K^+ ATPase (which pumps three Na^+ out, and two K^+ in, for every ATP consumed) and inward rectifying K^+ channels (which allows K^+ to flow out) return the electrocytes to their normal -85 mV resting potential. Since it is imperative that each electrocyte fires at the same time, a unique region of the brain called the pacemaker nucleus (or central control nucleus), located in the brain stem of electric fish, coordinates the simultaneous firing of the large population of electrocytes, and this unique combination of anatomical and biochemical specialization forms the basis for the electric organ discharge [5, 26].

Interspecies differences in electric organ signaling diversity have been explored most extensively in Gymnotiformes. Electric organ discharges vary between species due to each having a somewhat unique repertoire of other ion channels besides the ligand and voltage-gated Na+ channels [5]. Further, certain species of Gymnotiformes are electrically excitable on two membrane faces, which fire in succession and cause a unique waveform to the electric organ discharge. Also, the electric organ discharge between species can vary with the presence of 'accessory electric organs,' which are separated in space and activated at a slightly different time than the predominant electric organ in species; accessory electric organs do not generate their own distinct electric organ discharge, but rather, add complexity to the overall waveform of a given species electric organ discharge.

What is known about electric organs of weakly electric species?

Apart from the important historical role that strong-voltage electric fish such as *E. electricus* and the *Torpedo* played in the beginning days of modern physics and biochemistry, there are other important roles that weakly electric fish, particularly Gymnotiformes, have played in the natural sciences. Weakly electric fish are commonly used as models for the study of the evolution of intra-species communication, which confers species identity, and inter-species communication, or 'social' communication, which relays information about sexual identity (males and females often have differences in frequency of electric organ discharge), sexual receptiveness, and dominance [20].

Some research has gone into looking for signals that cause modulations to electric organ discharges [5]. Studies in Gymnotiformes looking at sexual dimorphism in electric organ signal have demonstrated that exposure to steroid hormones can alter the electric organ discharge over the course of days; for example, androgens have a masculinizing effect on the electric organ discharges of females [27, 28]. It was further demonstrated in both Mormyriformes and Gymnotiformes that androgens caused changes in electric organ discharges by acting on the electric organs themselves via changes in the duration of action potentials in electrocytes. This is achieved by modifying the kinetics of the voltage-gated Na+ channels, rather than by altering the rate of firing in the pacemaker nucleus of the brain [28, 29].

Some (but not all) Gymnotiformes exhibit social and circadian changes to electric organ discharge that occur much more rapidly (minutes/hours, as opposed to days with steroid hormones), and are under the control of peptide hormones [5]. In *Sternopygus,* it is

clear that these signaling changes, in part, are mediated by changes in membrane trafficking as circulating melanocortin peptide hormone caused very rapid changes in electric organ discharge via regulating ion channels trafficking to the electrocyte plasma membrane [30, 31]. Several other peptide hormones that are involved with regulating rapid changes in electric organ discharge have been described, and all of them are known now to signal through protein kinase A, but the phosphorylated targets of the kinase and the downstream players are currently not known [5]. No Mormyriformes have demonstrated rapid electric organ discharge modulation.

Importantly, Gymnotiformes exhibit one of the largest capacities for tissue regeneration, including the ability to regenerate parts of the brain, spinal cord, and parts of the body that contain diverse tissue types, such as all of the tissues lost from tail amputation [32]. A small number of studies have performed comprehensive analyses of the central nervous system and spinal cord in Gymnotiformes species, either under normal conditions or under repair conditions [33-35], but these mechanisms remain largely unexplained. Of particular interest, in *Sternopygus*, it has been demonstrated that post amputation of the claudal region of the tail, skeletal muscle transdifferentiates into electric organ during tail regeneration [36]. The underlying mechanisms of robust tissue regeneration in electric fishes are beginning to be characterized but are largely unexplained [32] and if electrical discharges are involved in this process, it will be of great interest and importance for translational work being performed in medical areas such as spinal cord injury.

What is known about the electric organs of *E. electricus?*

Like other Gymnotiformes, *E. electricus* has weak electric organ discharges that are used for the purposes of navigation and communication. *E. electricus* is unique among the Gymnotiformes, however, because it has the ability to produce strong voltage electric organ discharges for the purpose of predation and defense. Recently, it was shown that these strong-voltage discharges are not only used to stun prey, but to cause them to reveal their location by inducing muscle contractions, and thus, small voltages that are perceptible by the electroreceptive electric eel [37].

Very little is known about the differences among the three distinct electric organs in *E. electricus.* Electrocytes in the main electric organ are packed tightly next to one another while in the Sachs' electric organ, they are more separated from one another [38], which could contribute to the smaller voltage emitted from the Sachs' electric organ compared to the main electric organ. Additionally, electrocytes in the Sachs' electric organs are larger than the main electric organ and have larger invaginations [38]. Action potentials from main and Sachs' electric organ were pretty similar, but the action potentials occur much more rapidly in main electric organ—the mean duration being 1.5 msec in main and 2.2 msec in Sachs' [38]. Similar sizes of action potentials are observed in different electric organs, however [26, 38]. The energetic and regulatory demands on the electrocytes that are generating weak versus the strong electric organ discharges are quite different—the electrocytes generating weak electric organ discharges are experiencing action potentials continuously, while the electrocytes generating strong-voltage electric organ discharges are experiencing action potentials only intermittently. It is my hypothesis that the energetic demands on these organs is different, and that differences in the proteome will reflect this. Further, it my hypothesis that there may be post translational modifications on

the ion channels and pumps involved with the depolarization and repolarization of the plasma membrane and these will show differences among the organs.

A few molecular differences between the electric organs have been described, but almost entirely by measuring mRNA levels and electrophysiological measurements although the great abundance of the three major proteins (two sodium channels and the sodium pump) has enabled biochemical studies of these proteins for several decades. The highest expression of alpha subunit of Na+/K+ ATPase (ATP1a2a) is in the main electric organ and Hunter's electric organ and lowest expression is in the Sachs' electric organ [39]. With respect to the kinetics of the Na+/K+ ATPase, the main electric organ has highest V_{max} (how fast the enzyme performs its function) while Sachs' electric organ has the lowest, and Hunter's falls between the two [39]. Similarly, the Na+/K+ ATPase in main electric organ has lowest K_m (highest affinity) for Na^+ and K^+ of the three electric organs [39]. Together, this indicates that the main electric organ is more efficient at moving Na+/K+ out of the cell and can rapidly return to homeostasis [39].

A small number of studies have considered electric organ development in *E. electricus*. A study that gathered *E. electricus* right after hatching found that electrocytes are formed from electroblasts which develop from specific progenitor cells called pre-electroblasts. The pre-electroblasts are formed from the ventral most tip of embryonic trunk mesoderm at hatching [40]. Early in development actin filaments attached to Z-line like structures reminiscent of skeletal muscle are present in electric organ, and these disintegrate with time [40] as the cells lose their myogenic mechanical function and attain their enhanced electrogenic capabilities. The first electrocytes present in 10 mm hatchlings represent the beginning of the main electric organ which begins emitting electric organ discharges seven

days after hatching (15 mm in length). At 140 mm, the Hunter's electric organ can be detected, but the Sachs' electric organ is not yet present; interestingly, both low voltage and high voltage pulses are generated at this phase [25].

Problems in the field of electric fish biology

Electric fishes have played important roles throughout the sciences—historically, in understanding the nature of electricity, in the biochemical characterization of the first ion pumps and channels, in understanding the nature of the synapse and more modernly, in the behavioral and developmental biology sciences, in understanding the evolution of communication and understanding underlying mechanisms of tissue regeneration. Despite these important and diverse roles electric fishes have played in science, a comprehensive analysis of genes expressed in an electric organ of any species is lacking. Further, no genome sequence has been previously described for any of these species.

My dissertation thus provides the first analysis of the following important aspects of electric fish:

1. The annotated genome of the electric eel, *Electrophorus electricus* (Chapters 3, 4 and 5).
2. A comprehensive analysis of RNA and miRNAs expressed in eight tissues of E. electricus—whole brain, whole spinal cord, whole heart, whole kidney, skeletal muscle, main electric organ, Sachs' electric organ, and Hunter's electric organ (Chapter 3 and Chapter 4).
3. An analysis of mRNAs expressed in two other lineages of electric fishes with independently evolved electric organs (Chapter 4).

4. A quantitative proteomic and phosphoproteomic analysis of the three electric organs and skeletal muscle in *E. electricus* (Chapter 5).

In summary, this dissertation presents a series of studies that aim to shed light on the molecular nature and evolution of electric organs, both within *E. electricus* and also across independent lineages of electric fish.

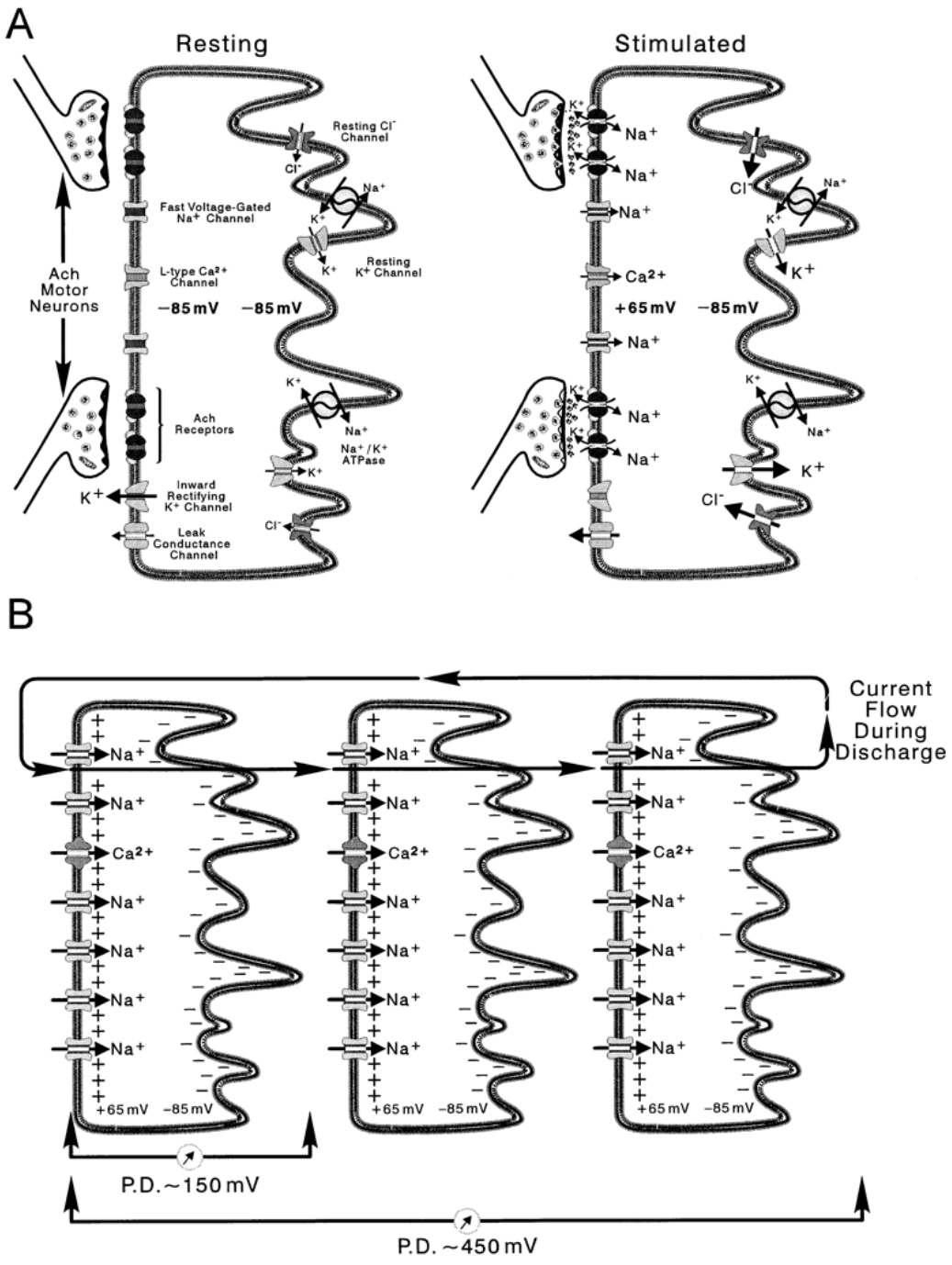


Figure 1.2. Image and caption from [26] "Diagrammatic representation of electrocytes. The left surface of each cell represents the posterior innervated membrane. (A) At rest, both the innervated and non-innervated membrane exhibit a potential of 85 mV. When stimulated, activated AchRs generate endplate potentials, triggering Na channel-mediated

action potentials peaking at 65 mV on the innervated membrane. The non-innervated membrane contains no voltage-gated Na channels and maintains the 85 mV resting potential. The result is a trans-cellular potential difference of approximately 150 mV. The presence of an L-type Ca^{2+} channel has not yet been supported by experimental evidence, but is included in this diagram given the myogenic origin of electric tissue. (B) Since each cell is stimulated simultaneously, electrocyte trans-cellular potentials summate. The potentials of three electrocytes culminate to produce 450 mV. Currents generated by stimulated electrocytes flow down electrocyte columns in the posterior to anterior direction. The circuit is closed by current flowing out the head of the eel, through the water, and back into the tail region."

References:

1. Moller P: A history of bioelectrogenesis. In: *Electric Fishes: History and Behavior.* London: Chapman & Hall; 1995.
2. Volta A, Banks J: On the electricity excited by the mere contact of conducting substances of different kinds. In a letter from Alexander Volta to Sir Joseph Banks, bart. n.p.,; 1800.
3. Finger S, Piccolino, M.: Epilogue. In: *The shocking history of electric fishes.* New York: Oxford University Press, Inc.; 2011.
4. Feldberg W, Fessard A: The cholinergic nature of the nerves to the electric organ of the Torpedo (Torpedo marmorata). *The Journal of physiology* 1942, 101(2):200-216.
5. Markham MR: Electrocyte physiology: 50 years later. *The Journal of experimental biology* 2013, 216(Pt 13):2451-2458.
6. Noda M, Takahashi H, Tanabe T, Toyosato M, Furutani Y, Hirose T, Asai M, Inayama S, Miyata T, Numa S: Primary structure of alpha-subunit precursor of Torpedo californica acetylcholine receptor deduced from cDNA sequence. *Nature* 1982, 299(5886):793-797.
7. Noda M, Shimizu S, Tanabe T, Takai T, Kayano T, Ikeda T, Takahashi H, Nakayama H, Kanaoka Y, Minamino N *et al*: Primary structure of Electrophorus electricus sodium channel deduced from cDNA sequence. *Nature* 1984, 312(5990):121-127.
8. Albers RW: Biochemical aspects of active transport. *Annual review of biochemistry* 1967, 36:727-756.
9. Bennett MV: Comparative physiology: electric organs. *Annual review of physiology* 1970, 32:471-528.

10. Bullock TH, Hopkins, C.D.: From electrogenesis to electroreception. In: *Electroreception.* Springer; 2005.
11. Zupanc GKHaB, T.H.: From electrogenesis to electroreception: an overview. In: *Electroreception.* Edited by Bullock TH, Hopkins, C.D., Arthur, P.N., Fay. R.R: Springer; 2005.
12. Bass AH: Electric organs revisited: Evolution of a vertebrate communication and orientation organ. In: *Electroreception.* Edited by Bullock TH, Heiligenberg, W.; 1986: 13-70.
13. Bennett MVL: Electric organs. *Fish Physiol* 1971, 5:347-491.
14. Moller P: Electric organs. In: *Electric Fishes: History and Behavior.* London: Chapman & Hall; 1995.
15. Mate SE, Brown KJ, Hoffman EP: Integrated genomics and proteomics of the Torpedo californica electric organ: concordance with the mammalian neuromuscular junction. *Skeletal muscle* 2011, 1(1):20.
16. Bray RN, Hixon MA: Night-Shocker: Predatory Behavior of the Pacific Electric Ray (Torpedo californica). *Science* 1978, 200(4339):333-334.
17. Nazarian J, Hathout Y, Vertes A, Hoffman EP: The proteome survey of an electricity-generating organ (Torpedo californica electric organ). *Proteomics* 2007, 7(4):617-627.
18. Koester DM: Anatomy and motor pathways of the electric organ of skates. *The anatomical record Part A, Discoveries in molecular, cellular, and evolutionary biology* 2003, 273(1):648-662.

19. Bennett MV, Grundfest H: The Electrophysiology of Electric Organs of Marine Electric Fishes : III. The electroplaques of the stargazer, Astroscopus y-graecum. *The Journal of general physiology* 1961, 44(4):819-843.
20. Albert JS, Crampton, W.G.R.: Electroreception and electrogenesis. Pp. 431-472 in The Physiology of Fishes, 3rd Edition. . In. Edited by Claiborne DHEaJB. Boca Raton, FL: CRC Press; 2005.
21. Baron VD: Electric discharges of two species of stargazers from the South China Sea (Uranoscopidae, Perciformes). *Journal of Ichthyology* 2009, 49(11):1065-1072.
22. White EG: The origin of the electric organs in Astroscopus guttatus. Washington, D.C.: Carnegie Institution of Washington; 1918.
23. Johnels AG: On the origin of the electric organ in *Malapterurus electricus*. *Quarterly Journal of Microscopial Science* 1956, 97(3):455-464.
24. Arnegard ME, Carlson BA: Electric organ discharge patterns during group hunting by a mormyrid fish. *Proceedings Biological sciences / The Royal Society* 2005, 272(1570):1305-1314.
25. Kirschbaum F, Schwassmann HO: Ontogeny and evolution of electric organs in gymnotiform fish. *Journal of physiology, Paris* 2008, 102(4-6):347-356.
26. Gotter AL, Kaetzel MA, Dedman JR: Electrophorus electricus as a model system for the study of membrane excitability. *Comparative biochemistry and physiology Part A, Molecular & integrative physiology* 1998, 119(1):225-241.
27. Mills A, Zakon HH: Coordination of Eod Frequency and Pulse Duration in a Weakly Electric Wave Fish - the Influence of Androgens. *J Comp Physiol A* 1987, 161(3):417-430.

28. Zakon HH, Dunlap KD: Sex steroids and communication signals in electric fish: a tale of two species. *Brain, behavior and evolution* 1999, 54(1):61-69.
29. Bass AH, Volman SF: From behavior to membranes: testosterone-induced changes in action potential duration in electric organs. *Proc Natl Acad Sci U S A* 1987, 84(24):9295-9298.
30. Markham MR, Allee SJ, Goldina A, Stoddard PK: Melanocortins regulate the electric waveforms of gymnotiform electric fish. *Hormones and behavior* 2009, 55(2):306-313.
31. Markham MR, McAnelly ML, Stoddard PK, Zakon HH: Circadian and social cues regulate ion channel trafficking. *PLoS biology* 2009, 7(9):e1000203.
32. Unguez GA: Electric fish: new insights into conserved processes of adult tissue regeneration. *The Journal of experimental biology* 2013, 216(Pt 13):2478-2486.
33. Zupanc MM, Wellbrock UM, Zupanc GK: Proteome analysis identifies novel protein candidates involved in regeneration of the cerebellum of teleost fish. *Proteomics* 2006, 6(2):677-696.
34. Salisbury JP, Sirbulescu RF, Moran BM, Auclair JR, Zupanc G, Agar JN: The central nervous system transcriptome of the weakly electric brown ghost knifefish (Apteronotus leptorhynchus): de novo assembly, annotation, and proteomics validation. *BMC genomics* 2015, 16(1):1354.
35. Sirbulescu RF, Zupanc GK: Spinal cord repair in regeneration-competent vertebrates: adult teleost fish as a model system. *Brain research reviews* 2011, 67(1-2):73-93.
36. Unguez GA, Zakon HH: Phenotypic conversion of distinct muscle fiber populations to electrocytes in a weakly electric fish. *The Journal of comparative neurology* 1998, 399(1):20-34.

37. Catania K: The shocking predatory strike of the electric eel. *Science* 2014, 346(6214):1231-1234.
38. Keynes RD, Martins-Ferreira H: Membrane potentials in the electroplates of the electric eel. *The Journal of physiology* 1953, 119(2-3):315-351.
39. Ching B, Woo JM, Hiong KC, Boo MV, Choo CY, Wong WP, Chew SF, Ip YK: Na+/K+-ATPase alpha-subunit (nkaalpha) Isoforms and Their mRNA Expression Levels, Overall Nkaalpha Protein Abundance, and Kinetic Properties of Nka in the Skeletal Muscle and Three Electric Organs of the Electric Eel, Electrophorus electricus. *PloS one* 2015, 10(3):e0118352.
40. Schwassmann HO, Assuncao MI, Kirschbaum F: Ontogeny of the electric organs in the electric eel, Electrophorus electricus: physiological, histological, and fine structural investigations. *Brain, behavior and evolution* 2014, 84(4):288-302.

Chapter 2: Overview of technologies used throughout this dissertation

Technological advances and increases in throughput in the fields of genomics, transcriptomics, and proteomics have provided researchers with the ability to realize the dynamics of biological systems, allowing for the exploration of not only the central dogma of molecular biology (genome, transcriptome, and proteome), but also elements that fall outside of the central dogma (microRNAs, protein phosphorylation, and others). Although the ability to generate massively high throughput datasets in biology has become commonplace, using and manipulating these valuable datasets as well as leveraging them to answer specific biological questions remains a technical challenge. This chapter provides an overview of the 'omics' technologies that I utilized throughout my dissertation work to answer biological questions related to electric fish, namely, genomics (Chapters 3, 4, and 5), transcriptomics (Chapters 3 and 4) and proteomics and phosphoproteomics (Chapter 5).

Genomics

The world of DNA sequencing changed in the 1970s with the advent of Sanger sequencing methods, and the subsequent automation of Sanger sequencing in the mid 1980s, which finally enabled scientists to sequence not only single genes with greater ease but also larger projects, such as the human genome [1-3]. The publically-funded Human Genome Project began in 1990, was budgeted for $3 billion, and anticipated to take 15 years. After some heavy competition, in 2001 two papers were released from the publically-funded Human Genome Project, and the privately owned Celera, showcasing the first Human Genome Assemblies [4, 5]. However, the method was costly and material (??) prohibitive and relatively low-throughput, and so the field of genome sequencing did not blossom until novel technologies were invented [3]. In the subsequent 14 years since these

papers were published, technology surrounding DNA sequencing has blossomed and the cost of sequencing has plummeted—one can now generate the sequence of short reads to attempt to assemble the sequence a large genome for ~$8000 [6].

The tides dramatically changed in the field of genomics in 2005 with the advent of new sequencing technologies, the very first instrument developed from 454 Life Sciences was capable of generating 50 times the sequence data at one-sixth the cause of traditional Sanger sequencing [3]. To date, several platforms and strategies exist; the choice of which strategies to use is a tradeoff among read length (for example, 100-250 base pairs with Illumina technologies, the potential of thousands of base pairs with Pacific Biosciences), accuracy of the sequencing reads generated (Pacific Biosciences reads have high error as high as ~13%, significantly higher than other sequencing platforms), cost, and number of unique reads generated [7].

One of the most ubiquitously used high throughput sequencing platform at this time is Illumina sequencing. Generally, sequencing methods involve isolating genomic DNA and randomly breaking it into smaller fragments (< 500 bp), followed by ligation of platform-specific adapters, amplification of fragments, followed by sequencing [7]. While the user can generate a massive amount of small fragment data with relatively small cost, these small genomic fragments (up to 300 bp for Illumina technology) offer very little in terms of 'genomic context.' For genomes that are highly repetitive, short sequence fragments are limited, because in large stretches of repetitive sequence, short reads generated are not unique and can match multiple places, making assembly across the region impossible. Mate-pair libraries can be advantageous because they can allow users to bypass stretches of repetitive sequence in genomes, which stymie assemblies by short reads alone. Mate pair

libraries allow users to sequence the ends of much longer genomic fragments (for example, 5kb). Although the user does not generate sequence data across the entire fragment (only the ends are sequenced, and the reads generated themselves are still short reads), the user gains 'genome context' information because it is known that each short read in a mated-pair are separated by the distance that is your library size (so ~5kb in this example). These kinds of reads are valuable when assembling a large genome from short reads, because they allow connections to be made between contiguous DNA pieces in the assembly (contigs, which are assembled regions of consensus sequence), and allowing the genome to be scaffolded into longer, ordered pieces.

A technology that has the potential to radically change the field of genomics is Pacific Bioscience (PacBio), which has the ability to sequence single molecules of DNA and generate extremely long read lengths; a current major limitation of this technology, however, is the error rate, remains quite high (~13%) [8]. Despite this high error rate, longer read information is incredibly useful in genome assembly because it offers more genomic context for the assembler to build off of, and clever programmers have created methods to correct these error-prone PacBio reads, including using short Illumina reads, aligning them to PacBio, and using the low error rate Illumina reads to correct errors in the long PacBio reads [9, 10].

Genome assembly

Necessitated by the rapid expanse of sequencing projects, the number of genome assembly programs available and the strategies that go along with them have expanded. Researchers can benefit from the rapidly expanding knowledge of genome assemblers—many assembly programs have been evaluated and rigorously tested by those in the field

for performance on real and simulated data such as the Assemblathon and GAGE collaborations [11, 12]. In these large-scale collaborations, experts in genome assembly worked together to explore how different assemblers and assembly parameters affect the assembly outcome on the same raw sequence input; this was done massively, for small bacterial genomes and large vertebrate genomes alike. These kinds of projects are incredibly useful for scientists learning genome assembly because it gives them a way to choose not only the programs that may work best for their purposes but also the ingredients for the 'recipes' of genome sequence (amount and combinations of read lengths, etc.) and assemblers that worked best for similar genomes.

There are many strategies that exist in genome assembly. A common strategy utilizes an abundance of both short reads (eg. 100 bp reads from on an Illumina platform) and reads that are generated from a longer-range paired end library (2x 100 bp reads from a library with 5 kb or 12kb insert sizes, for example). As was performed in our first round of genome assembly (Chapter 3, 4), short reads may be assembled together to form contiguous regions of assembled DNA (called contigs) using primarily short-read (2x100) and short-jump (2.5 kb library) mate pair data [13]. This assembly was further improved by performing an additional round of scaffolding with longer-range mate pair data (12kb insert library, in this case), which attempted to string the assembled contigs into larger pieces, which give more genome context (Chapter 5). Following scaffolding, short reads can be used to attempt and fill in gaps within the resulting scaffolds. This reflects the strategy ultimately used within my dissertation, but many pipelines and algorithms exist and have been successful, and generally several attempts are needed to find a strategy that will work well on any given dataset [14].

Because so many genome assembly programs exist, and because no assembly is perfect, evaluating and comparing genome assemblies is very important. Various metrics help researchers compare assemblies to one another, and these commonly include: number of contigs (or scaffolds) in the assembly; size (in nucleotides) of the longest contig (or scaffold); the N50 length, which is the length of the longest contig (or scaffold) for which 50% of the total nucleotides in your assembly are in contigs greater than, or equal to [15]. Unfortunately, a comprehensive approach at comparing genome assembly versions remains elusive (for example, longer scaffolds in an assembly may not mean those scaffolds are of higher accuracy). In my evaluation of genome assemblies, I used the statistics discussed above, but also considered the underlying genes predicted in the assemblies when determining whether I had made an improvement to a given genome assembly (see next section).

Gene model prediction and annotation

Generally, a genome assembly itself is not particularly useful or informative by itself. It is only when the information encoded within the genome is deciphered that the genome becomes more readily useful. Just as many genome assemblers exist, many programs and pipelines exist for the annotation of large, vertebrate genomes (AUGUSTUS, and MAKER are examples of these [16, 17]). Gene prediction programs benefit from the incorporation of additional input that can give the prediction algorithm hints as to what gene models should look like. For example, gene prediction programs can use hints in the form of RNA sequencing reads mapped to the genome, or protein sequence from related species mapped to the genome, as evidence of the location of genes as well as the exon/intron boundaries of

those genes. Further, several gene prediction programs can be trained to predict genes in novel genomes; this process requires many hundreds to thousands of well-defined gene models to be known and used as "training" and "testing" sets, by which the gene prediction program iteratively predicts genes and compares them to the known gene models as to make slight adjustments to the parameters and improve gene prediction as a whole for a given genome.

For the work presented in this dissertation, I used AUGUSTUS to do all of the gene predictions in both genome assemblies (see Chapters 4 and 5). Although as part of the annotation pipeline MAKER can utilize AUGUSTUS to perform gene prediction, MAKER doesn't allow the user to make many parameter changes to the way AUGUSTUS performs. In my hands, running AUGUSTUS on a small subset of scaffolds with slight parameter changes to the way AUGUSTUS uses extrinsic information (RNA seq information, etc) and the Human Genome preset parameters, comparing the outcomes, and choosing the extrinsic file parameters that give the best results outperforms what I could produce from MAKER. Further, running AUGUSTUS with the Human Genome preset parameters instead of training AUGUSTUS for the *E. electricus* genome gave the best overall outcome of genes predicted. I used this strategy in gene prediction in Chapters 4 and 5.

To date, the process of assembling and annotating a large genome requires some technical knowhow—at a minimum, the ability to operate in a Linux environment, and likely, the ability to write scripts (ie. perl or python language); the need for these skills and the lack of them in biology is so prevalent that free introductory courses are available and relatively easy to find [18]. Genome-scale projects often require analysis that are far from "push button," but projects such as Galaxy have come a long way towards making

otherwise intensive bioinformatics analyses more accessible to researchers by consolidating commonly used bioinformatics tools, providing data formatting ability, establishing pipelines for data analysis, and improving ease of reproducibility and data sharing [19-21].

Transcriptomics

RNA sequencing to measure transcript expression

The high throughput measurement of all transcripts in a sample (called transcriptomics) is often facilitated by RNA sequencing, another application of high throughput sequencing used to profile all transcripts in a sample (such as tissue type) [22]. In RNA sequencing, total RNA is isolated and converted to cDNA with reverse transcriptase. Often, if only messenger RNA sequencing is desired, RNAs are enriched by the presence of the poly-A tail. Alternately, if only micro RNAs are desired, only short cDNAs will be size-selected and used. Following this, the cDNA libraries are prepared and sequenced using high-throughput technologies. The reads that are generated are then mapped to a reference genome or reference transcriptome for quantitation; the number of reads mapping to any given transcript or gene is proportional to the abundance of that transcript in the sample being tested. Many tools exist to perform the steps of this process, such as mapping reads to your genome or transcriptome [23-26], to count the reads mapping to individual genes in your genome, or transcripts in your assembly [25, 27], and to normalize resulting abundances, and perform differential expression analysis [28]; this list is non-exhaustive, and is continuously growing, but these are the programs utilized in this dissertation (in Chapters 4, and 5).

Transcriptome sequencing and assembly

Just like genomes, transcriptomes (a collection of all mRNA transcripts present in a sample) can be assembled *de novo* from sequencing reads. Several programs exist to perform transcriptome assembly, and depending on the assembler used, there is a tradeoff between completeness of the transcriptome assembly (Trinity maximizes transcripts reported) and reliability (Abyss only reports highly reliable assemblies) [29]. The transcriptome assemblies of five species of electric fishes are presented in Chapter 4, and for each assembly, the program Trinity was used.

Transcriptome sequencing and assembly hails enormous advantages for investigators, especially in organisms for which there is no genome sequence. Using transcriptome assembly programs, one can gain a comprehensive understanding of gene expression in an organism or tissue as well as yield a predicted proteome of good quality. There are still gains to be had from transcriptome assembly, however. It is my experience that transcriptome assembly, while comprehensive, can yield an unmanageable number of assembled transcripts (hundreds of thousands to millions, see supplemental materials in Chapter 4). Depending on the scientific question, this may be problematic—it is my experience that it is incredibly difficult or impossible to sort out orthologous sequences (especially on a large-scale) among fishes when dealing with hundreds of thousands of sequences, hundreds of which can be variants of a single 'gene' that have incorporated low frequency transcriptional events or errors in them. The presence of the *E. electricus* genome in our study presented in Chapter 4, with a limited number of gene models predicted in the genome (~20,000 gene models instead of ~400,000 assembled

transcripts), was a major facilitator of comparisons among species, because we had a manageable anchor to which to compare all other species. It would be interesting to revisit our transcriptome assemblies, given the field has had a few additional years to improve since these original assemblies were completed, and to try reassembling them using programs that weren't available or weren't fully matured at the time we were doing these initial experiments (Chapter 4). In fact, if further experiments were to be done using the transcriptome assemblies alone, I would be a strong advocate to spending additional time and using new resources to improve the assemblies because improving the transcriptome assemblies to get a smaller number of higher-confidence transcripts would improve and simplify the downstream analysis. Although Trinity generally continues to outperform other *de novo* transcriptome assemblers [30, 31], in addition to improvements in transcriptome assembly programs, new tools to compare the output from several assemblies have been developed [31]. By using such tools to evaluate transcriptome assemblies quickly and objectively, researchers could spend additional time tweaking assembly parameters and quickly and objectively evaluating output.

Proteomics

Untargeted, "shotgun" proteomics approaches

Mass spectrometers are instruments used in proteomics to precisely measure the mass of proteins and/or short peptides derived from them. Untargeted proteomics is an unbiased method that can be used to identify thousands of proteins in a complex sample such as plant or animal tissue; by contrast, targeted proteomics approaches are used to distinctly target and monitor small number of proteins. Untargeted proteomics approaches

typically follow a similar workflow [32]: proteins are extracted from cells or tissues of interest. Extracted proteins are digested enzymatically into peptides with enzymes such as trypsin or lys-c. Following digestion, complex samples can be greatly simplified using prefractionation methods like ion-exchange chromatography or high pH reversed-phase chromatography. Samples can also be enriched for low-abundance post-translational modifications, such as phosphorylation or acetylation. Prepared peptide samples can then be analyzed by liquid-chromatography coupled with tandem mass spectrometry (also called MS/MS or MS_2).

Database development, searching, and false discovery rate calculation

Large-scale proteomics experiments, such as those described in Chapter 5, are not *de novo* peptide sequencing, and rather, rely heavily on a complete reference proteome database, although methods are being developed for proteomics analysis without a reference proteome [32]. To date, for peptide sequence identification, a well-defined proteome and a database search algorithm are required to bioinformatically correlate the thousands of experimentally obtained MS_1 and MS_2 peptide spectra with the theoretically-derived fragmentation spectra from a protein database to determine peptide and protein identifications. For that reason, I spent additional time in the course of my thesis work to improve the genome assembly, over what we used initially for RNA sequencing analysis (see "genome assembly" section above). Once a high quality protein database has been developed, various database search algorithms exist for this purpose, both free (such as OMSSA [33]) and fee-for-license (such as MASCOT (Matrix Science)). Because the output from database searching is the highest-scoring match between the experimentally obtained

mass spectra, and the theoretical peptide spectra generated in a protein database, and thus, the two are intimately intertwined, it is important to distinguish between true and false peptide matches. One approach at identifying and filtering out false identifications (false discovery rates, FDR), uses a decoy database that is generated by taking the reverse of all protein sequences in your actual protein database, giving the database search algorithm lots of opportunities to match spectra to peptide sequences that should not exist in the sample you are measuring [34]. While analyzing data post database searching, an FDR threshold is chosen which strikes a balance between being identifying the maximal number of peptides or proteins in the sample and being highly confident that all peptides/proteins identified are true matches.

Protein Quantitation

Untargeted proteomics approaches can be quantitative and multiplexed with the addition of isobaric mass tags. One such approach, utilizing Tandem Mass Tags (TMT, Thermo Scientific) allows for the addition of unique isobaric tags to peptide samples post digestion. Labeled peptides from independent biological samples can then be combined in equal ratios and analyzed simultaneously on a mass spectrometer (at the time of this writing, a TMT kit to maximally label a 10-plex reaction was available). The distinct isobaric tags add identical mass/charge ratios to the labeled peptides; in the first mass scan (MS_1) identical peptides from different biological samples are concurrently selected. During subsequent fragmentation, the isobaric tags release a reporter ion moiety along with the peptide, which are then detected in MS_2; the relative abundance of each reporter ion determines the relative abundance of each peptide in each sample.

Phosphoproteomics

Mass spectrometry is a widely used tool for examining post-translational modifications of proteins such as phosphorylation–the addition of a phosphate group to serene, threonine, or tyrosine. Phosphorylation is known to play an important role in modulating the activity of many proteins, including ion channels and pumps [36, 37]. This is of particular interest with respect to electric organ discharge (EOD) in electric fishes; because EODs in electric fishes are caused by ion fluctuations within electrocytes, it follows that phosphorylation of the key ion channels and pumps likely plays an important role in the EODs of electric fish. Since the number of peptides containing even one phosphorylated amino acid is low, in examining a tryptic digest of a proteome, it is necessary to enrich for phosphopeptides prior to injection onto the MS/MS instrument. The method that I used to enrich for phosphopeptides is titanium dioxide chromatography [35], one of several widely used in this field.

Summary

By leveraging these different high throughput approaches spanning genomics, transcriptomics, and proteomics, we were able to comprehensively characterize electric organs in electric fishes for the first time. First, by sequencing, assembling, and annotating the genome of *E. electricus*, and transcriptome assemblies of several other electric fish species, we were able to compare abundances of transcripts expressed in EO and muscle across lineages, ultimately leading to hypotheses regarding both the development and evolution of electric organs in distinct lineages of electric fish species (Chapter 4). The

genome of *E. electricus* and the discrete set of proteins predicted in it, facilitated proteomic and phosphoproteomic comparisons among the three distinct EO, leading to the identification of novel phosphorylation sites in proteins important for electric organ discharge, and the development of hypotheses related to ATP conservation in the continually-used weak EO (Sachs' EO) and intermittently-used strong-voltage organ (main EO) (Chapter 5). Together, this dissertation provides a reference genome, several transcriptomes, and a proteome/phosphoproteome that will provide a foundation for the electric fish community for decades to come.

References:

1. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M: Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 1977, 265(5596):687-695.
2. Sanger F, Nicklen S, Coulson AR: DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* 1977, 74(12):5463-5467.
3. Schuster SC: Next-generation sequencing transforms today's biology. *Nature methods* 2008, 5(1):16-18.
4. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al*: The sequence of the human genome. *Science* 2001, 291(5507):1304-1351.

5. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*: Initial sequencing and analysis of the human genome. *Nature* 2001, 409(6822):860-921.

6. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) [http://www.genome.gov/sequencingcosts.]

7. Metzker ML: Sequencing technologies - the next generation. *Nature reviews Genetics* 2010, 11(1):31-46.

8. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y: A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics* 2012, 13:341.

9. Au KF, Underwood JG, Lee L, Wong WH: Improving PacBio long read accuracy by short read alignment. *PloS one* 2012, 7(10):e46679.

10. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED *et al*: Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature biotechnology* 2012, 30(7):693-700.

11. Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R *et al*: Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience* 2013, 2(1):10.

12. Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M *et al*: GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome research* 2012, 22(3):557-567.

13. Gallant JR, Traeger LL, Volkening JD, Moffett H, Chen PH, Novina CD, Phillips GN, Jr., Anand R, Wells GB, Pinch M *et al*: Nonhuman genetics. Genomic basis for the convergent evolution of electric organs. *Science* 2014, 344(6191):1522-1525.

14. Baker M: De novo genome assembly: what every biologist should know. *Nature methods* 2012, 9(4):333-337.

15. Gurevich A, Saveliev V, Vyahhi N, Tesler G: QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013, 29(8):1072-1075.

16. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B: AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids research* 2006, 34(Web Server issue):W435-439.

17. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M: MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome research* 2008, 18(1):188-196.

18. Unix and perl primer for biologists [http://korflab.ucdavis.edu/Unix_and_Perl/index.html]

19. Goecks J, Nekrutenko A, Taylor J, Galaxy T: Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology* 2010, 11(8):R86.

20. Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J: Galaxy: a web-based genome analysis tool for experimentalists. *Current protocols in molecular biology / edited by Frederick M Ausubel [et al]* 2010, Chapter 19:Unit 19 10 11-21.

21. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J *et al*: Galaxy: a platform for interactive large-scale genome analysis. *Genome research* 2005, 15(10):1451-1455.

22. Wang Z, Gerstein M, Snyder M: RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews Genetics* 2009, 10(1):57-63.

23. 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.

24. Langmead B, Trapnell C, Pop M, Salzberg SL: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* 2009, 10(3):R25.

25. Li B, Dewey CN: RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* 2011, 12:323.

26. Trapnell C, Pachter L, Salzberg SL: TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009, 25(9):1105-1111.

27. Anders S, Pyl PT, Huber W: HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 2015, 31(2):166-169.

28. Anders S, Huber W: Differential expression analysis for sequence count data. *Genome biology* 2010, 11(10):R106.

29. Nagarajan N, Pop M: Sequence assembly demystified. *Nature reviews Genetics* 2013, 14(3):157-167.

30. Chopra R, Burow G, Farmer A, Mudge J, Simpson CE, Burow MD: Comparisons of de novo transcriptome assemblers in diploid and polyploid species using peanut (Arachis spp.) RNA-Seq data. *PloS one* 2014, 9(12):e115055.

31. Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R, Dewey CN: Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome biology* 2014, 15(12):553.

32. Altelaar AF, Munoz J, Heck AJ: Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature reviews Genetics* 2013, 14(1):35-48.

33. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH: Open mass spectrometry search algorithm. *Journal of proteome research* 2004, 3(5):958-964.

34. Elias JE, Gygi SP: Target-decoy search strategy for mass spectrometry-based proteomics. *Methods in molecular biology* 2010, 604:55-71.

35. Sugiyama N, Masuda T, Shinoda K, Nakamura A, Tomita M, Ishihama Y: Phosphopeptide enrichment by aliphatic hydroxy acid-modified metal oxide chromatography for nano-LC-MS/MS in proteomics applications. *Molecular & cellular proteomics : MCP* 2007, 6(6):1103-1109.

36. Levitan IB: Modulation of ion channels by protein phosphorylation and dephosphorylation. *Annual review of physiology* 1994, 56:193-212.

37. Poulsen H, Morth P, Egebjerg J, Nissen P: Phosphorylation of the Na+,K+-ATPase and the H+,K+-ATPase. *FEBS letters* 2010, 584(12):2589-2595.

Chapter 3: Unique patterns of transcript and miRNA expression in the South American strong voltage electric eel (*Electrophorus electricus*).

The work presented in this chapter has been published:

Traeger, L.L[*]., Volkening, J.D[*]., Moffett, H., Gallant, J.R., Chen, P-H., Novina, C.D., Phillips, G.N. Jr., Anand, R., Wells, G.B., Pinch, M., Guth, R., Unguez, G.A., Albert, J.S., Zakon, H[+]., Sussman, M.R[+]., Samanta, M.P[+]. (2015). BMC Genomics. ######

- [*] indicates lead author; [+] indicates corresponding author

*** All supplementary files are available in the published version of this manuscript, online.

Abstract

Background

With its unique ability to produce high-voltage electric discharges in excess of 600 volts, the South American strong voltage electric eel (*Electrophorus electricus*) has played an important role in the history of science. Remarkably little is understood about the molecular nature of its electric organs.

Results

We present an in-depth analysis of the genome of *E. electricus*, including the transcriptomes of eight mature tissues: brain, spinal cord, kidney, heart, skeletal muscle, Sachs' electric organ, main electric organ, and Hunter's electric organ. A gene set enrichment analysis based on gene ontology reveals enriched functions in all three electric organs related to transmembrane transport, androgen binding, and signaling. This study also represents the first analysis of miRNA in electric fish. It identified a number of miRNAs displaying electric organ-specific expression patterns, including one novel miRNA highly over-expressed in all three electric organs of *E. electricus*. All three electric organ tissues also express three conserved miRNAs that have been reported to inhibit muscle development in mammals, suggesting that miRNA-dependent regulation of gene expression might play an important role in specifying an electric organ identity from its muscle precursor. These miRNA data were supported using another complete miRNA profile from muscle and electric organ tissues of a second gymnotiform species.

Conclusions

Our work on the *E. electricus* genome and eight tissue-specific gene expression profiles will greatly facilitate future research on determining the coding and regulatory

sequences that specify the function, development, and evolution of electric organs. Moreover, these data and future studies will be informed by the first comprehensive analysis of miRNA expression in an electric fish presented here.

Background

The electric eel (*Electrophorus electricus*) is a freshwater teleost (order: Gymnotiformes) from South America, the only species identified to date within the genus *Electrophorus* [1]. Reaching more than seven feet in total length, *E. electricus* is most famous for its ability to generate strong voltage discharges (up to ~600 volts [2]) from electric organ (EO) tissues for use in predation and defense. Because of this remarkable ability, *E. electricus* has played a prominent role in the history of science – in physics, for early insights into the nature of electricity, and in biochemistry, as a rich source of tissue for extensive biochemical investigations of ion channels and pumps [3].

Over 700 species of electric fishes have been identified [1], the vast majority of which are capable of generating only weak electric organ discharges (EOD) for the purpose of navigation and communication. Like other members of Gymnotiformes, *E. electricus* produces weak EODs (mV-V scale) for navigation and communication. However, it is unique within Gymnotiformes in possessing three distinct EOs (most other Gymnotiformes have only one distinct organ, and some have additional accessory organs), and it is the only gymnotiform capable of a strong-voltage EOD. In *E. electricus*, strong EODs are produced from the main EO and the anterior two-thirds of the more ventrally-positioned Hunter's EO; weak EODs are produced from the Sachs' organ and the posterior one-third of the Hunter's organ (Figure 3.1). All three EOs of *E. electricus* are derived developmentally from

a germinal zone located on the ventral margin of the hypaxial musculature [4,5]. Interestingly, the ability to produce EODs is not limited to the Gymnotiformes; indeed, electric organs have evolved independently from skeletal muscle at least six times in fishes [4,6].

Despite the importance of electric fishes in the history of science, genomic, proteomic, and metabolomic approaches towards understanding the molecular nature of electrocytes (the single cell of the electric organ) have only lately been undertaken [7-16]. In a recent report from this consortium, we described a small group of protein-coding genes that showed similar patterns of expression in electric organs as compared to skeletal muscle from three distinct lineages in which the electrogenic phenotype evolved independently [13]. Included in this study was the first draft genome sequence of *E. electricus*, but a detailed analysis of gene content and tissue-specific expression in this electric fish species remained to be described. In this report, we describe the first comprehensive analysis of genes and multi-organ gene expression of *E. electricus*. Our gene set enrichment analysis using Gene Ontology terms found that genes highly expressed in EOs are enriched in functions pertaining to transmembrane transport, receptor signaling, and hormone binding. We performed the first analysis of microRNA (miRNA) expression in an electric fish and show that all three EOs in *E. electricus* express a unique repertoire of miRNAs, including a novel miRNA and three conserved miRNAs involved with muscle development inhibition in mammals. The results build a framework for comprehensively understanding the molecular nature of an electrocyte and provide a foundation for future work on electric organs in electric fish.

Results

E. electricus genome features

We used next-generation sequencing technologies to sequence and assemble the genome of *E. electricus* and the transcriptome of the three EOs and five other tissues: brain, spinal cord, heart, skeletal muscle, and kidney, as described previously [13] (Table 3.S1a). A set of 29,363 gene models representing an estimated 22,000 protein-coding genes was annotated from the genome and transcriptome. Comparison between the genomes of *E. electricus* and *Danio rerio*, the nearest related sequenced fish genome, showed considerable local synteny (i.e., *hox* genes, see Figure 3.S1a). The average intron size in *E. electricus* was similar to that of the other sequenced non-pufferfish teleosts and was ca. one-third that of *D. rerio* (Table 3.S2).

E. electricus transcriptome analysis

Our comparison of genes expressed in eight organs of *E. electricus* [13] showed that the mRNA expression profiles of electrocytes found in the three EOs (Hunter's, Sachs' and main) were distinct from all other cell types, with a greater similarity in gene expression to skeletal and heart muscle as compared to kidney, brain or spinal cord (Figure 3.2). This finding was consistent with the known myogenic origin of electrocytes in most species [4]. Variance filtering of the gene models predicted in our first computational annotation removed ~ 3/4 of the genes with low covariance among tissues. A subsequent k-means clustering (k = 12) revealed sets of tissue-specific co-transcriptionally regulated genes ([13] and Figure 3.3). Of particular interest were clusters 1, 6, 7, 9, and 10, which represented genes over-expressed only in EOs (cluster 9), genes over-expressed in skeletal and heart muscle (cluster 1), genes over-expressed in both skeletal muscle and EO (cluster

6), genes over-expressed in skeletal muscle, heart and EO (cluster 7), and genes over-expressed in brain, spinal cord and EO (cluster 10) (Figure 3.3). Clusters 6 and 7 represented a shared identity between electrocytes and myocytes, while clusters 1 and 9 represented sets of genes that were down- or up-regulated in electrocytes compared to myocytes and may hold clues to the unique structure and function of the EO.

In order to understand what functions were enriched and best characterized in our tissue-specific expression clusters, we performed a Gene Ontology (GO) enrichment analysis on each of the 12 tissue-specific clusters. This analysis revealed enriched functions that were consistent with expectations based on the tissues in the tissue-specific clusters (Additional file 2). For example, GABA-A receptor activity, ionotropic glutamate receptor activity, and extracellular-glutamate-gated ion channel activity appear enriched in clusters 3 and 4, both of which are gene clusters over-expressed in brain and spinal cord. In cluster 11 (kidney), enriched GO terms are consistent with fish kidney (fish kidneys are used not only for osmoregulation but also for hematopoiesis), including several GO terms involved with transmembrane transport as well as heme binding.

GO analysis of cluster 9 (all EOs) showed an enrichment of GO terms involved with transmembrane transporting (Figure 3.4), while enriched GO terms of cluster 1 (skeletal muscle and heart) consisted of calcium-transporting ATPase activity, voltage-gated calcium channel activity, and calcium ion binding, highlighting the well-known role of Ca^{2+} in muscle contraction. In cluster 6 (all EOs and skeletal muscle), the most enriched GO terms involved the general category of transcriptional regulation, including sequence-specific DNA binding, ligand-activated sequence-specific DNA binding, and sequence-specific DNA binding transcription factor activity, as well as GO terms involved with acetylcholine

receptor activity. In cluster 7 (all EOs, skeletal muscle, and heart), the enriched GO terms were involved in metabolism, such as NADP binding, NAD (P) + transhydrogenase activity, and phosphofructokinase activity, as well as insulin-like growth factor (IGF) I and II binding. In cluster 10 (all EOs, brain, and spinal cord), the enriched GO terms interestingly included voltage-gated Ca^{2+} activity and Ca^{2+} binding. Additionally, cluster 10 shows enrichment in receptor binding, receptor activity, and hyaluronic acid binding.

TopGO was further used to generate GO graphs for each of the five primary clusters of interest (1, 6, 7, 9, & 10) using enriched GO terms (p-value < 0.05) as input (Figure 3.S2). The findings were consistent with those of Figure 3.4, but additionally highlight highly represented categories unique to each cluster. Especially informative are the GO graphs generated for clusters 6, 7, and 9. The GO graph generated for the enriched GO terms in cluster 6 (skeletal muscle and electrocytes) has a large, highly represented group broadly characterized as metabolism (7 out of 19 total terms). Additionally, broad categories including actin/tropomyosin binding are also over-represented in cluster 6; this is intriguing, as the electrocytes have lost their contractile machinery. Also, skeletal muscle cells and electrocytes are activated by acetylcholine, and this is reflected in the graph, including the GO terms for acetylcholine-activated cation-selective channel activity and acetylcholine receptor activity.

Similar to the results from cluster 6, the GO graph generated for the enriched terms in cluster 7 (skeletal muscle, heart, and all EOs) had a highly represented, broad category of metabolism (7 out of 20 total terms). Finally, the cluster 9 (all EOs) GO graph has a highly represented group characterized as transmembrane transport (7 of 16 total terms). This group includes GO terms such as voltage-gated sodium channel activity and inward

rectifier K^+ channel activity, which are directly involved in electric organ discharge (EOD). Cluster 9 also includes GO terms involved in hormone/androgen binding; this GO term is physiologically relevant as EOD has been shown to be regulated in part by the presence of sex hormones [17].

Hoxc cluster expression in EO

One surprise arising from the 8-tissue profiling was the elevated expression of *hoxc10a*, *hoxc11a*, *hoxc12a* and *hoxc13a* genes in all three electric organs (Figure 3.S1b). Hox family members are well-known components of the regulatory machinery that specifies the anterior-posterior body axis of animals, and in many cases the spatial expression patterns of *hox* genes within tandem modules on the genome have been observed to correlate with spatial distribution of expression along that axis [18]. We observed the same set of *hox* genes (*hoxc10a*, *hoxc11a*, *hoxc12a*, *hoxc13a*) from the *hoxca* cluster to be over-expressed in two other Gymnotiformes (*Eigenmannia virescens* and to a lesser extent *Sternopygus macrurus*) and one mormyrid (*Brienomyrus brachyistius*) as well (Figure 3.S1c and [13]); interestingly, these *hoxc* genes are not highly expressed in the EO of the electric catfish *Malapterurus electricus* (data not shown). Jawed vertebrates have four paralogous *hox* cluster genes (*hoxa*, *hoxb*, *hoxc*, *hoxd*), among which only the *hoxc* cluster was shown to be dispensable for body plan development. The entire cluster was lost in Elasmobranch fishes and its deletion in mice caused only minor transformations of axial identity [19-22]. Whether these are retained in adults to specify the predominant posterior location of the electrocytes in Gymnotiformes and mormyrids (but not electric catfish) or have another function is not known. Our observation raises the possibility of neofunctionalization of posterior *hoxca* genes in some species of electric fishes.

Analysis of binding sites for highly upregulated transcription factors in EO

A significant future goal is to understand the mechanisms underlying EO development and maintenance. As a step toward that goal, a plausible hypothesis is that transcription factors highly upregulated in EO regulate distinctive characteristics of EO. Based on this hypothesis, candidates for this set of important transcription factors were identified by their high expression ratios (>7-fold) in EO compared to skeletal muscle (Additional file 3, A). Within this set, a subset of transcription factors were particularly promising candidates because they were also highly expressed in EO compared to all five non-EO tissues including skeletal muscle: *egr3* (early growth response 3), *six2a* (sine oculis-related homeobox 2a), *hoxc11a* (homeo box C11a), *foxj3* (forkhead box J3), *ar* (androgen receptor), *pou3f1* (POU class 3 homeobox 1), *lbx2* (ladybird homeobox homolog 2), and *hoxc10a* (homeo box C10a).

One possible explanation for how cluster 9 genes become upregulated in EO is enrichment of binding sites within their promoters for one or more transcription factors that are themselves highly expressed in EO. To test this hypothesis, we examined putative promoter regions within cluster 9 genes for binding sites of 21 transcription factors highly expressed in EO relative to skeletal muscle, using 2984 randomly-sampled genes as a background control (Additional file 3). Their DNA binding sites were frequently found in putative promoter regions of cluster 9 genes (see Additional file 3, B for examples). From testing the density of binding sites in promoters of all cluster 9 genes as a group compared with background controls, the smallest p-values suggesting binding site enrichment were found for transcription factors *prrx1b* (paired related homeobox 1b; $p = 0.006$) and *lbx2* (ladybird homeobox homolog 2; $p = 0.049$) (Additional file 3, C). From the subset of 51 of

the cluster 9 genes most highly expressed in EO relative to skeletal muscle (Additional file 3, D), the smallest p-values suggesting enrichment were found for transcription factors *prrx1b* ($p = 0.02$) and *emx2* (empty spiracles homeobox 2; $p = 0.047$). We then compared the number of occurrences of each transcription factor binding site in the promoters of the 51 most highly expressed cluster 9 genes individually against our background control and found p-values smaller than 0.005 for *emx2*, *lbx2*, *pou3f1* (POU class 3 homeobox 1), *prrx1b*, and *sox12* (SRY-box 12) (Additional file 3, E). These results suggest particular highly upregulated transcription factors that might contribute to upregulation of cluster 9 genes in EO through selective enrichment of their DNA binding sites and are possible targets for further study. It is important to note, however, that none of these p-values remained significant at a 5% FDR after multiple testing correction, and that the magnitude of change compared to background for the genes/binding sites discussed was generally relatively small.

Parallel evolution in the Kir2 channel and the Na+/K + −ATPase

It has been reported that some electrocyte-specific ion channels involved in generating the electric discharge appear to be evolving at a higher than expected rate in electric fishes (see [23,24] for a discussion of the $Na_V1.4a$ sodium channels in electrocytes). Two teleost-specific members (*kcnj2b*, *kcnj12b*) of the inward rectifying K^+ channel (Kir2) family are abundant in *E. electricus* electrocytes. A hallmark of Kir2 channels is a highly conserved aspartate residue at the inner mouth of the channel that binds Mg^{2+} and polyamines and plugs the channel at depolarizing voltages imparting rectification [25,26]. In the non-rectifying members of the Kir family, there is an asparagine residue instead at that site. Within the channels encoded by the gymnotiform *kcnj2b* and *kcnj12b*, both have

an asparagine at the Mg^{2+} binding site, suggesting that the gymnotiform electrocyte has a unique intracellular environment. In addition, the α2 isoform of the sodium pump, which is highly over-expressed in the electrocyte, shows an amino acid substitution at a conserved site (Figure 3.S3). In an interesting case of parallel evolution, the same substitution occurs in squid, although there it is due to RNA editing rather than a permanent change in the codon. This amino acid change is thought to enhance sodium transport [27].

Reduced vision and loss of opsin genes in *E. electricus*

Diurnally active teleost fishes generally have four physiologically distinct cone types in their retinae; one with long, one with medium, and two with short wavelength-sensitive opsins [28]. In contrast, *E. electricus* is nocturnally active and often lives in muddy rivers and streams where the ambient light is strong in longer wavelengths [29,30]. We searched the *E. electricus* genome for opsin genes; interestingly, we found only long (red), and medium (green) but no short-wave sensitive (blue and violet) cone opsin genes (Figure 3.S4). Although possible, it is unlikely that this pattern is due to incomplete sequencing coverage of the genome as we also recovered a gene for the rod photopigment rhodopsin and numerous other teleost non-photopigment opsins such as melanopsin. We hypothesized that the lack of short-wave sensitive cone opsin genes may be shared with other species that live in similarly muddy and murky conditions. Indeed, when we probed an EST database of a species in a sister group of Gymnotiformes, a non-electrogenic catfish [31], we observed the same cone opsin profile as *E. electricus*.

MiRNA analysis: novel sequences and EO-specific expression

MiRNAs are an evolutionarily ancient class of small non-coding RNAs that regulate many gene networks during animal development [32]. MiRNA composition and expression

levels have been used as a molecular taxonomy approach for categorization of tissue types, description of cellular physiological states and even classification of disease states [33]. It has also been suggested that expansion of miRNA families has played a central role in the remarkable morphological complexity among vertebrates [34]. To investigate the potential role of miRNAs in electrocyte phenotype and function, we isolated and sequenced small RNAs from the spinal cord, brain, heart, skeletal muscle, kidney, and all three EOs of *E. electricus*. We identified 294 conserved miRNAs belonging to 119 miRNA families expressed in one or more of the eight tissues [35-38]. We also identified 18 novel miRNAs from the set of unmatched reads with perfect matches to the *E. electricus* genome (Figure 3.5c). As shown in other organisms, conserved miRNAs tend to be more robustly expressed than species-specific miRNAs [39-41] (Figure 3.5b). However, all novel miRNAs found in *E. electricus* showed tissue-specific expression patterns, suggesting that they may serve specific functions in *E. electricus*.

To investigate the role of miRNAs in the EO, we performed hierarchical clustering of miRNA expression across the eight *E. electricus* tissues. MiRNA expression patterns clustered nervous tissue from the brain and spinal cord separately from cardiac muscle, skeletal muscle, and EOs, and clustered EOs more closely with skeletal muscle than cardiac muscle (Figure 3.5a). Indeed, from the 313 total miRNAs identified (294 conserved and 18 novel) only 18 showed high differential expression between skeletal muscle and the three EOs (Figure 3.6a). Conserved miRNAs with lower expression in EO than in muscles have annotated roles in muscle development and differentiation in other organisms [42-44]. Three of these miRNAs are muscle-specific, also called “myomiRs” (miR-133, miR-206, and miR-499), because they play critical roles in muscle development and function. In contrast,

miRNAs with higher expression in EO than in muscle include multiple miRNAs with roles including inhibition of muscle differentiation (miR-193, miR-218, and miR-365) [45,46].

The most upregulated miRNA in EO compared to muscle is the novel miRNA mir-11054 (Figure 3.6a). This electrocyte specific "electromiR" was 30-fold higher in all three EOs compared to skeletal muscle and was not detected in brain, spinal cord, kidney, or heart tissues of *E. electricus*. Notably, mir-11054 is expressed from a locus that is important to electrocyte function, an intron of the inward-rectifier K^+ channel gene *kcnj12b* that is abundant in electrocytes. Though it is expressed from a region overlapping the *kcnj12b* gene, mir-11054 is expressed in the antisense direction from an intronic sequence that is unique to *E. electricus* (Figure 3.6b). This "electromiR" mir-11054 has no known homologue in any fish or mammal to date.

To further probe the importance of the differentially expressed miRNAs in EO function, we sequenced miRNA libraries from the EO and skeletal muscle of a second electric fish, *S. macrurus*. This comparison revealed interesting differences and similarities between the two Gymnotiformes species (Figure 3.6a). ElectromiR mir-11054, which is highly transcribed in *E. electricus* EO from the K+ channel intron, is not detected in *S. macrurus*, suggesting that it may be specific to *E. electricus*. In addition, our results indicate that, in contrast to the downregulation observed in *E. electricus* EO, *S. macrurus* EO expresses most myo-miRs at levels similar to those found in skeletal muscle. The continued expression of myomiRs in *S. macrurus* EO is consistent with a more 'muscle-like' phenotype in this fish [47]. In contrast to myomiRs, miRNAs with annotated roles in inhibiting muscle differentiation are commonly upregulated in the EO of *E. electricus* and *S. macrurus*. These miRNAs include miR-218a, which inhibits cardiac muscle development [48], and the

bicistronically-encoded miR-365a and miR-193a, which inhibit skeletal muscle development [49].

Discussion

The analyses presented here describe molecular comparison of *E. electricus* electric organs to muscle and other tissues and build off of this consortium's previous work exploring the convergent evolution of electric organs in independent fish lineages [13]. Analyses of the *E. electricus* genome identified a number of interesting characteristics. The *E. electricus* genome is approximately ~700 Mb in size, which is roughly half of that of *D. rerio*. By comparing gene models across available fish genomes, we found that *E. electricus* intron lengths were about one third that of *D. rerio* (Table 3.S2), which likely is a significant contributing factor to the difference in genome sizes among the sequenced non-pufferfish teleosts and *D. rerio*. We also found a number of genomic changes that contribute to the adaptation of *E. electricus* environment and physiology. For example, within the α2 subunit of the sodium pump, which is highly abundant in EO, there is an amino acid substitution that has been demonstrated previously to occur in squid and is thought to enhance sodium transport [27] (Figure 3.S3). Given the important role of sodium transport in EOD, this substitution may be important for electrocytes to rapidly relieve high internal [Na^+] after depolarization. Of interest, this finding seems to be specific to *E. electricus*, as *E. virescens* and *S. macrurus* do not share this substitution. As another example, *E. electricus* lives in muddy rivers and streams where short wavelength light is more easily filtered out [29,30]. We were unable to identify short-wave light sensitive opsin genes in the *E. electricus* genome despite being able to identify long and medium-wave opsin genes. Upon

examination of the non-electrogenic catfish that resides in a similar environment [31], we observed a similar opsin profile. Thus, the loss of short wavelength-sensitive opsins is likely adaptive as it allows for a greater number of photoreceptors with opsins in the most useful portions of the visual spectrum. The absence of short wavelength-sensitive opsins has also been reported in most mammals living under dim light conditions or which, like bats and cetaceans, utilize other specialized sensory systems (e.g. echolocation) [50,51]. It is interesting to note that catfish took a different path in adapting to nocturnal living by developing taste buds all over their body to 'taste' the environment [52], much as electric fish sense their surroundings using electroreceptors.

One main goal of our research was to characterize the genes expressed in each of our eight tissues, with particular focus on understanding the unique repertoire of genes expressed in EO compared to skeletal muscle. To that end, we clustered our genes by similarity of expression (Figure 3.3), and within the resulting clusters employed GO term enrichment techniques to characterize enriched functions (Figures 4 and S2). The EO-specific cluster (cluster 9) showed unique aspects of EO that are not shared with muscle (the tissue type from which it is developmentally derived). Within this cluster, we found an abundance of genes with "transmembrane transporting" function and a lack of "cytoskeletal binding" and other contraction-related terms. This result reflected the tradeoff between contractility-related physiological function and increased electrogenic output. Conversely, we identified an enrichment of metabolism-related genes within all myogenic tissues (skeletal muscle, heart, and EO), which suggested that the basic metabolic processes in muscle are retained in electrocytes from myogenic precursors. Recent efforts from this consortium have identified IGF signaling as an important element in the

independent evolution of electric organs [13]. Interestingly, our functional enrichment analysis revealed enrichment for IGF binding in myogenic tissues (skeletal muscle, heart, and EO). Looking into this further, although IGF binding was enriched both in muscles and EO of *E. electricus* in the analyses presented here, only the *regulators* of this signaling pathway were up-regulated in electric organs compared to skeletal muscle in *E. electricus*, in other Gymnotiformes, and in two other lineages of fishes with independently evolved electric organs [13], indicating IGF signaling likely has a specialized role in EO over muscle. As we previously reported [13], our hypothesis is that IGF signaling contributes to the increase in cell size of electrocytes over muscle fibers.

Within the cluster of genes co-expressed in EO, brain, and spinal cord (cluster 10) we found enrichment for terms pertaining to Ca^{2+} transport and binding. This may highlight a gain of an additional Ca^{2+} function in EO that utilizes genes expressed typically in brain and spinal cord despite no need for Ca^{2+} for contraction (indeed, EO has turned down expression of genes relating to Ca^{2+} function). However, we cannot rule out the possibility that nerve contamination in EO tissue (which is highly innervated) contributed to this result, a possibility that would require additional experimentation to rule out.

A crucial step in our greater quest to elucidate the mechanisms by which electric organs have evolved is understanding the underlying principles of how genes are regulated in EO compared to muscle. Given that several transcription factors are highly upregulated in EO, we hypothesized that there may be enrichment of these transcription factor binding sites in the promoters of genes that are also highly expressed in EO. However, our attempts at identifying enriched binding motifs within the promoters of genes highly expressed in EO (cluster 9 genes) failed to identify a “smoking gun”; thus, it is reasonable to suspect that

other mechanisms beyond enrichment of DNA binding sites for highly expressed transcription factors in EO might be responsible for regulation of cluster 9 genes. First, binding sites for transcription factors that are not upregulated might be enriched in cluster 9 genes and lead to transcriptional upregulation in EO. This mechanism, however, does not explain the restriction of cluster 9 gene upregulation to only EO. As an additional mechanism, factors regulating gene transcription, including chromatin state, might differentially affect availability of these binding sites in cluster 9 genes compared to availability in genes of other clusters. The upregulation of seven transcription factor genes (*hoxc10a*, *hoxc11a*, *hoxc12a*, *hoxc13a*, *six2a*, *sox11b*, and *mef2b*) in EO from four electric fishes (*E. electricus*, *S. macrurus*, *E. virescens*, and *B. brachyistius*) [13] suggests their particular importance in EO identity. The lack of enrichment of binding sites for these seven, however, suggests that other mechanisms beyond transcription factor upregulation are involved in expression of cluster 9 genes—an exciting area for future research studies.

MiRNAs play important roles in regulating gene networks throughout animal development [32]. We aimed to characterize miRNA expression in our eight tissues of interest, with particular focus on muscle and EO and with the goal to determine whether there was a potential role of miRNA in EO development or maintenance. Our analysis revealed nearly 300 conserved miRNAs with known functions and 18 novel miRNAs; these novel miRNAs showed tissue-specific expression patterns which indicated they may be serving tissue-specific functions in *E. electricus*. Of particular interest were the 18 miRNAs that showed high differential expression among skeletal muscle and three EOs (Figure 3.6a), and, to gauge whether our findings were *E. electricus*-specific or shared among Gymnotiformes, we also performed miRNA sequencing and expression analysis on *S.*

macrurus. Of particular note were three conserved muscle-specific miRNAs that were highly expressed in EO relative to skeletal muscle (miR-193, miR-218 and miR-365) in both Gymnotiformes tested; these miRNA have known roles in inhibiting muscle differentiation [45,46]. Interestingly, we identified a novel "electromiR" abundantly expressed in *E. electricus* EO but not identified in *S. macrurus* EO, implying that this novel miRNA arose for *E. electricus*-specific electrocyte development and function (Figure 3.6b). The upregulation of conserved miRNAs with known roles in blocking muscle development in the EO of both *E. electricus* and *S. macrurus* provides evidence that miRNAs are part of a common toolkit involved in the development and maintenance of the electrocyte phenotype in Gymnotiformes. Uncovering the functional role of miRNAs that are uniquely expressed in *E. electrophorus* EOs could shed light on the molecular mechanisms involved in the modification of the muscle program to give rise to such a specialized tissue as the EO.

Conclusions

We describe here an analysis of the first sequenced genome of an electric fish (*E. electricus*) and of mRNA and miRNA libraries from eight organs including the three electric organs. This study, which builds upon previous work from this group focusing on shared protein-coding gene expression patterns between EO and skeletal muscle in multiple independent lineages of electric fish [13], provides a focused and thorough examination of both novel genomic characteristics as well as protein- and microRNA-encoding gene expression patterns and gene set enrichment in a panel of diverse organs from the strong voltage electric eel. Genes expressed in electric organs were enriched for functions involving transmembrane transport, whereas skeletal muscle showed enrichment for

contraction-related functions, reflecting the specialization of electrocytes for electrogenesis over contraction. Gene expression shared between skeletal muscle and electric organs had functional enrichment for genes relating to metabolism, suggesting that metabolic characteristics of each cell type are similar even though the chemical energy is transduced to a different degree in terms of mechanical versus electrical energy. The first comprehensive analysis of miRNA expression in electric fish identified three conserved miRNAs that have known roles in inhibiting muscle development as highly expressed in electrocytes, suggesting that miRNAs may be playing an important role in electrocyte development and maintenance. Interestingly, one of the 18 novel miRNAs identified was highly specific to EO and was transcribed from the reverse strand of an intron within the potassium channel that is also highly expressed in EO. Future studies will build from the work presented here to understand more deeply the function and evolution of genes expressed in electric organs, including the molecular and evolutionary distinction of the strong voltage electric organ unique to *E. electricus* in the Gymnotiformes lineage.

Methods

Analysis of gene ontology enrichment in clusters with tissue-specific expression

E. electricus gene models that were described previously [13] as being short fragments of whole genes and labeled “split” or “split_scaff” were filtered out for this analysis. Gene Ontology (GO) terms assigned to each *D. rerio* gene were downloaded from Ensembl (release version 71). GO terms were mapped back to all possible *E. electricus* genes using their *D. rerio* assignments. A total of 13,647 *E. electricus* genes remained in the analysis after these steps. In order to prevent the presence of a single GO term in one of our

gene clusters from coming up as significant, we filtered out all GO terms from the tissue-specific clusters that were present less than two times. Enriched GO terms in each of the 12 tissue-specific clusters were identified by using the "elim" method of topGO [53], with a minimum node size of 3. The elim method of topGO attempts to eliminate some of the local dependencies inherent to the GO graph structure by removing genes mapped to higher level GO terms, as to emphasize lower level (more specific) GO terms. We used the Fisher's exact test to identify significantly enriched GO terms with a p-value < 0.01. For clusters 1, 6, 7, 9, and 10, we used topGO to create the GO graph resulting from these enriched terms.

MiRNA sequencing and analysis

We measured miRNA expression in eight tissues (brain, spinal cord, heart, skeletal muscle, Sachs' electric organ, main electric organ, Hunter's electric organ, and kidney) of *E. electricus*.

Small RNA library preparation and sequencing in E. electricus

One (1) μg total RNA from each tissue (isolated as described above for mRNA sequencing) was used to prepare small RNA sequencing libraries using an Illumina TruSeq Small RNA Library Preparation Kit according to the manufacturer's instructions (Illumina, Inc., San Diego, CA) with one modification: cDNA was size-selected on an agarose gel in the 140–300 bp range. Each of seven tissues (brain, spinal cord, heart, muscle, Sachs', main, and Hunter's EO) was labeled with a unique index using PCR indexing primers. The indexed small RNA libraries were pooled and sequenced on a single lane of HiSeq (1x100bp) at the same time as the corresponding mRNA was sequenced in our previous study ([13]). The kidney small RNA followed an identical protocol, except that it was done at a later date and

was not indexed to be pooled with other samples. The linker sequences were removed from all libraries, and sets of trimmed reads of length 20, 21, 22 and 23 were compiled.

Identification of conserved and novel miRNA genes in E. electricus

Conserved *E. electricus* miRNA orthologs were identified by BLASTN comparison of expressed sequences from small RNA libraries to sequences from the Rfam database [54]. A set of potential novel miRNAs were generated by removing matches to conserved miRNAs, as well as contaminating small RNAs matching to a database of rRNA, snoRNAs, snRNAs, and mitochondrial rRNA and tRNAs derived from *Danio rerio*, *Tetraodon nigroviridis*, and *Takifugu rubripes* sequences in the Ensembl database [55]. To identify genomic precursors for conserved and novel miRNAs, sequences were aligned to the *E. electricus* genome, and perfect matches together with 140 bp of flanking sequence were retrieved. Secondary RNA structure was predicted using RNAFold from the Vienna package [56], and pre-miRNA hairpin structures were identified by the methods of [35]. MiRNAs were named according to their ortholog, and novel miRNAs were given unique names.

Expression profiling and clustering

For expression analysis, small RNA libraries were matched to conserved and novel hairpin precursors in *E. electricus*, using a relaxed standard which allowed up to two base pair mismatches between sequences to account for known processes which add non-template nucleotides to the 3' end of miRNAs [57], and for sequencing errors. Expression of 3p and 5p products was identified and calculated separately. Small RNA reads were adapter trimmed and filtered to 20–23 nucleotides in length. From this, 93% of small RNA reads from *E. electricus* tissue matched conserved miRNA families found in the *E. electricus* genome. Small miRNA sequences from *S. macrurus* were matched to precursors from *E.*

electricus because of the lack of availability of a *S. macrurus* genome. MiRNA read counts for each tissue profiled in *E. electricus* and *S. macrurus* were first normalized by correction for the median number of total reads, followed by linear normalization (Additional file 4). MiRNAs which did not have at least 16 normalized reads in at least one tissue were discarded from further analysis. Complete linkage clustering of tissue-specific miRNA expression using the Pearson correlation distance (Figure 3.5a) was performed using Genepattern [58].

Analysis of transcription factor binding sites in cluster 9 genes

Searching the putative promoter sequences with MatInspector (Genomatix) with default settings for core and matrix similarity identified potential binding sites for transcription factors in promoter regions. A promoter was defined as the region 2 kilobases upstream from the transcription start site of a gene as determined by AUGUSTUS gene modeling. For statistical analysis, the number of binding sites for a transcription factor in putative promoters of cluster 9 genes as a group or as single genes was compared to the expected number. The expected number was derived from the binding sites in putative promoters of a set of 2984 randomly selected genes from all clusters. The p-value of a comparison was defined as the cumulative probability from a binomial distribution of the expected number of binding sites greater than or equal to the number of binding sites identified with MatInspector in cluster 9 genes. The p-values reported were not corrected for the number of comparisons.

The DNA binding properties of 21 highly upregulated *E. electricus* transcription factors selected for analysis with MatInspector were assumed to be identical to the binding properties of the homologous vertebrate transcription factors as described in MatBase

(Genomatix). The relevance of DNA binding properties in MatBase to the properties of these *E. electricus* transcription factors has not been experimentally determined. Amino acid identities greater than 88% and averaging 96% for the DNA binding domains in these *E. electricus* transcription factors compared with mouse homologs supports the use of the MatInspector database for the analysis.

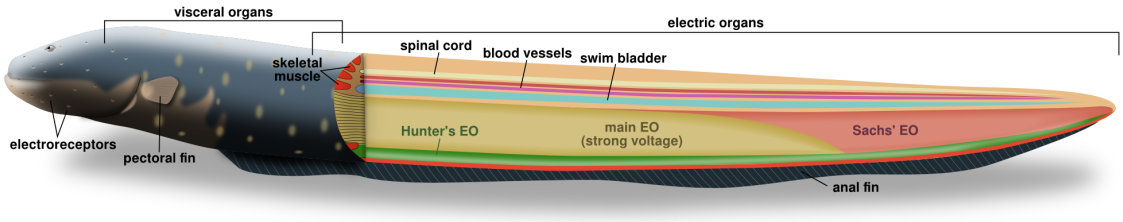


Figure 3.1 Overview of electric eel anatomy. Longitudinal section of *E. electricus* showing location and relative size of the three electric organs along with other anatomical features.

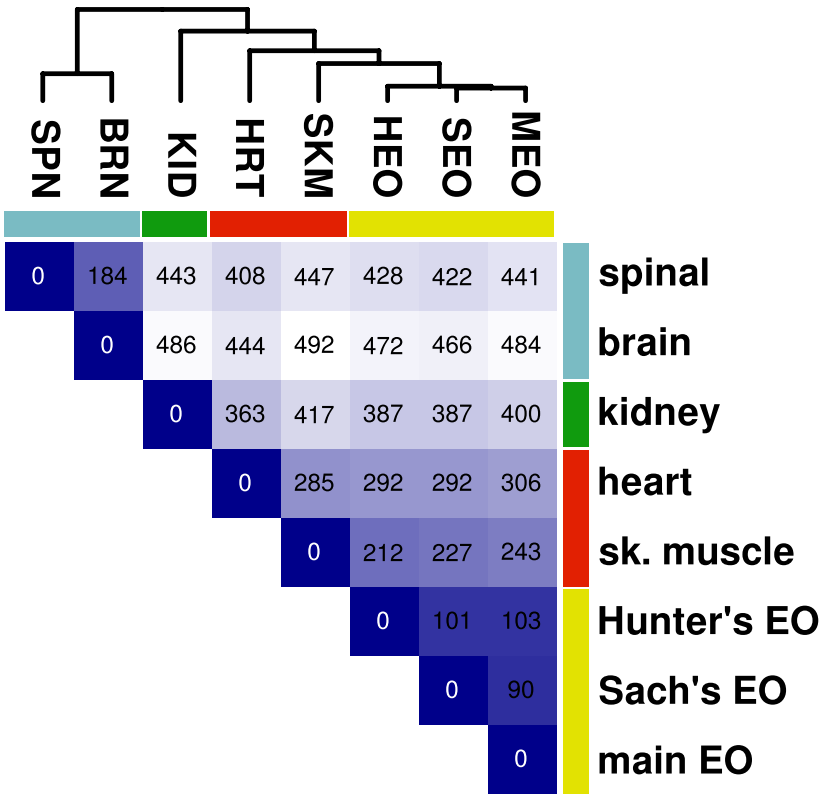


Figure 3.2 Clustering of eight electric eel tissues by gene expression profile. Gene expression values for the eight tissues were normalized, variance filtered, $\log_2$-transformed and median-centered as described previously [13]. Values shown are Euclidean distances based on ca. 6,000 genes passing the covariance filter, also indicated by blue shade (darker indicates shorter distance). Clustering was performed using complete linkage hierarchical clustering. Colored bars indicate a general grouping by tissue and cell type that is suggested by the data, with electric organ tissues (yellow) clustering most closely with skeletal and heart muscle (red). SPN = spinal cord; BRN = brain; KID = kidney; HRT = heart; SKM = skeletal muscle; HEO = Hunter's EO; SEO = Sachs' EO; MEO = main EO.

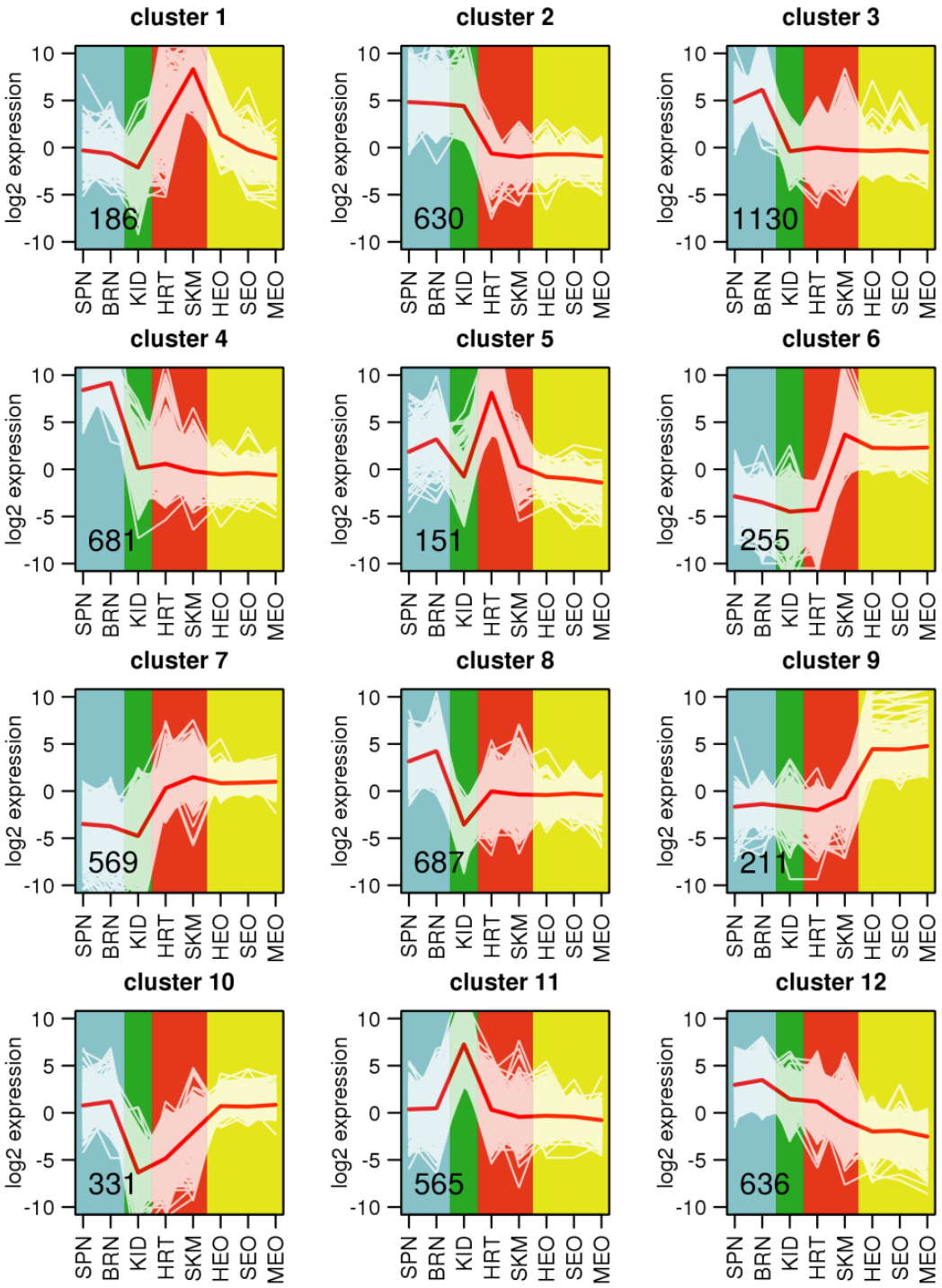


Figure 3.3 Clustering of co-expressed genes in *E. electricus*. Figure reproduced from [13]. A k-means clustering analysis (k = 12) was performed as previously described [13]. Values in lower-left indicate the number of genes in each cluster. White plot lines represent

log_2-transformed and median-centered expression of individual genes and red plot lines show median values for the cluster. Background shading indicates general categories of tissue/cell type. SPN = spinal cord; BRN = brain; KID = kidney; HRT = heart; SKM = skeletal muscle; HEO = Hunter's EO; SEO = Sachs' EO; MEO = main EO.

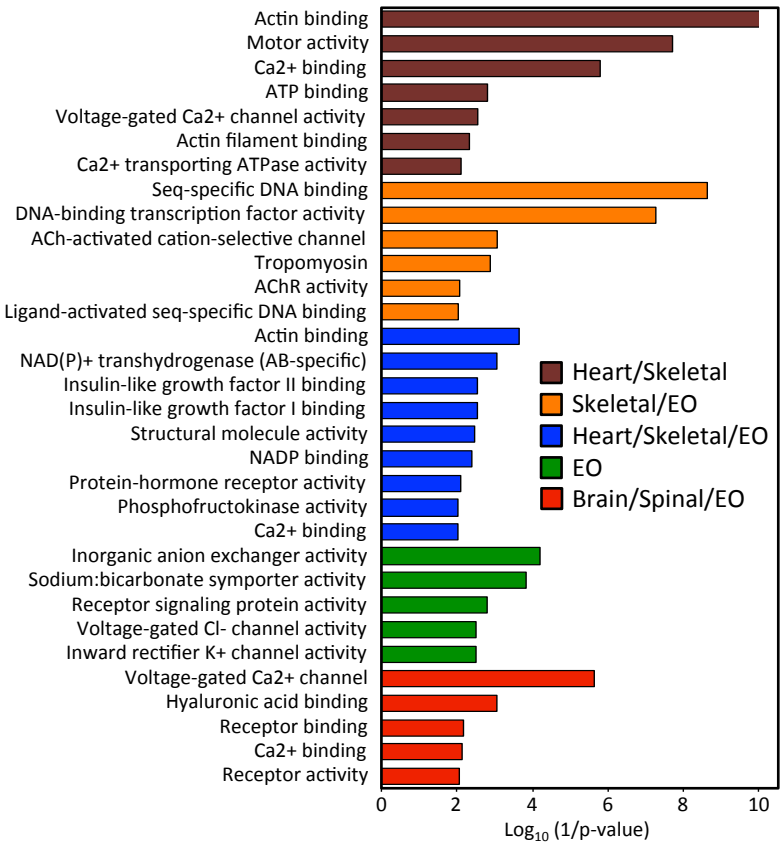
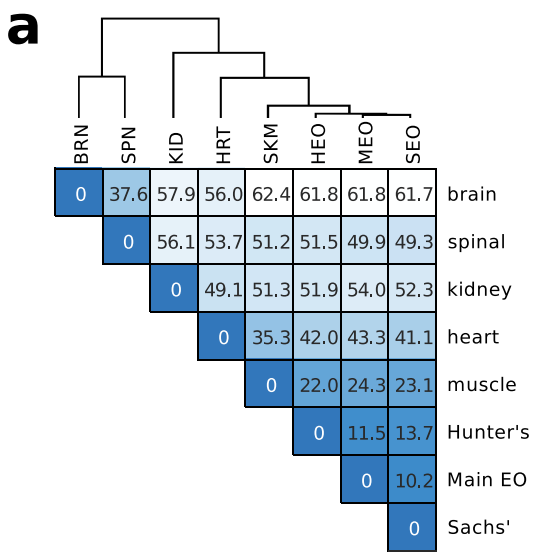
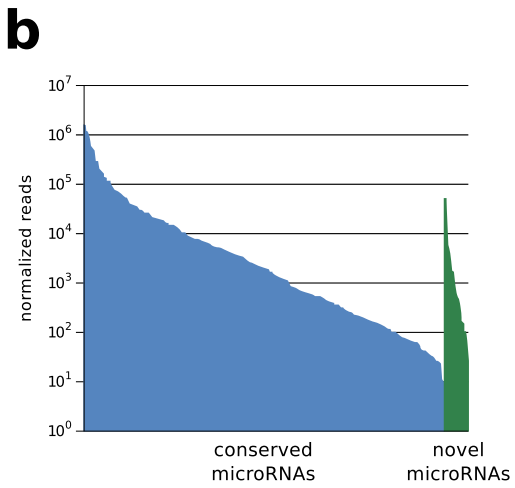


Figure 3.4 Gene Ontology enrichment of genes over-expressed in muscle and electric organ of *E. electricus*. Enrichment of GO terms in the "molecular function" ontology. Shown are enriched GO terms identified using topGO in cluster 1 (over-expressed in skeletal and heart muscle), 6 (over-expressed in skeletal muscle and EO), 7 (over-expressed in skeletal muscle, heart and EO), 9 (over-expressed only in EOs), and 10 (over-expressed in brain, spinal cord and EO) ($p < 0.01$).


a
BRN
SPN
KID
HRT
SKM
HEO
MEO
SEO
0 37.6 57.9 56.0 62.4 61.8 61.8 61.7 brain
0 56.1 53.7 51.2 51.5 49.9 49.3 spinal
0 49.1 51.3 51.9 54.0 52.3 kidney
0 35.3 42.0 43.3 41.1 heart
0 22.0 24.3 23.1 muscle
0 11.5 13.7 Hunter's
0 10.2 Main EO
0 Sachs'


b
normalized reads
10^7
10^6
10^5
10^4
10^3
10^2
10^1
10^0
conserved microRNAs
novel microRNAs

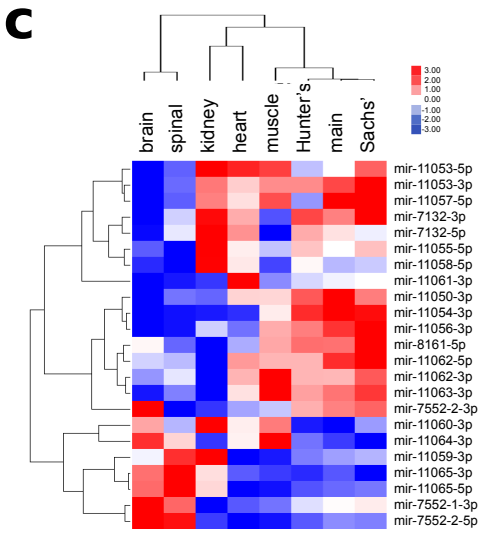

c
brain
spinal
kidney
heart
muscle
Hunter's.
main
Sachs'
3.00
2.00
1.00
0.00
-1.00
-2.00
-3.00
mir-11053-5p
mir-11053-3p
mir-11057-5p
mir-7132-3p
mir-7132-5p
mir-11055-5p
mir-11058-5p
mir-11061-3p
mir-11050-3p
mir-11054-3p
mir-11056-3p
mir-8161-5p
mir-11062-5p
mir-11062-3p
mir-11063-3p
mir-7552-2-3p
mir-11060-3p
mir-11064-3p
mir-11059-3p
mir-11065-3p
mir-11065-5p
mir-7552-1-3p
mir-7552-2-5p

Figure 3.5 Known and novel miRNA genes. (**a**) MiRNA expression classifies *E. electricus* tissues. Tissue distance matrix based on miRNA expression. MiRNA expression values for 8 tissues were filtered, normalized and $\log_2$ transformed as described in Methods. Values shown are Euclidean distances. Tree is derived from complete linkage hierarchical clustering. (**b**) Normalized sequencing read counts for conserved and novel *E. electricus* miRNAs. (**c**) Heatmap and complete linkage hierarchical clustering of novel miRNA $\log_2$-transformed and median-centered expression values in *E. electricus* tissues demonstrates tissue-specific expression patterns.

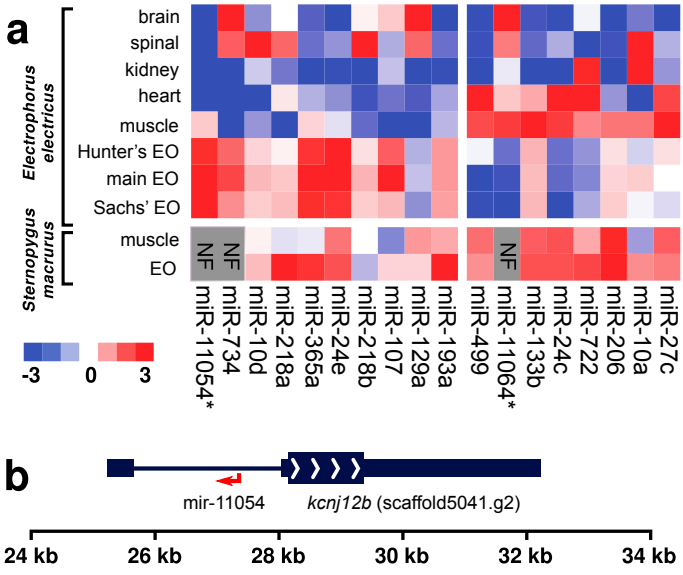


Figure 3.6 Electrocyte-specific microRNA expression. (**a**) Heatmap of miRNA expression in *E. electricus* and *S. macrurus.* Values are $\log_2$-transformed and median-centered values of tissue-specific expression for each miRNA, such that blue indicates under-expression and red indicates over-expression relative to the median. The miRNAs shown are limited to those with >4-fold increased or decreased expression in *E. electricus* electric organs compared to skeletal muscle. Log_2 values are clamped between −3 and +3. Asterisks indicate novel *E. electricus* miRNAs. (**b**) Schematic diagram of the *kcnj12b* gene locus and novel electro-miR mir-11054 on scaffold5041 of the *E. electricus* genome. Thin boxes are UTRs, the thick box with white directional arrows is the coding sequence, and the thin line is an intron. The red arrow on the antisense strand indicates the location of the novel mir-11054 microRNA.

Figure 3.S1

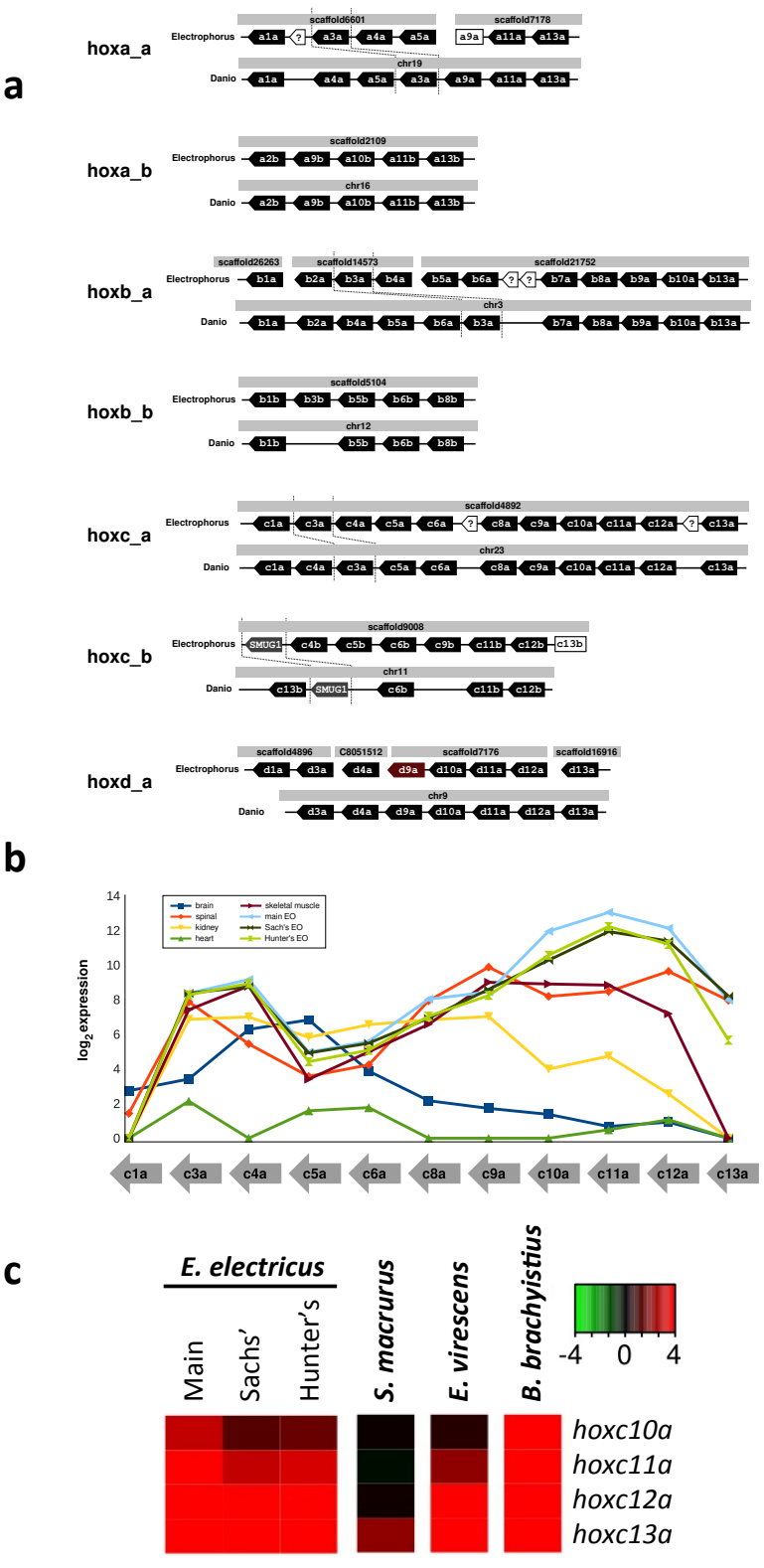

a
hoxa_a
Electrophorus
scaffold6601
scaffold7178
a1a ? a3a a4a a5a a9a a11a a13a
Danio
chr19
a1a a4a a5a a3a a9a a11a a13a
hoxa_b
Electrophorus
scaffold2109
a2b a9b a10b a11b a13b
Danio
chr16
a2b a9b a10b a11b a13b
hoxb_a
Electrophorus
scaffold26263
scaffold14573
scaffold21752
b1a b2a b3a b4a b5a b6a ? ? b7a b8a b9a b10a b13a
Danio
chr3
b1a b2a b4a b5a b6a b3a b7a b8a b9a b10a b13a
hoxb_b
Electrophorus
scaffold5104
b1b b3b b5b b6b b8b
Danio
chr12
b1b b5b b6b b8b
hoxc_a
Electrophorus
scaffold4892
c1a c3a c4a c5a c6a ? c8a c9a c10a c11a c12a ? c13a
Danio
chr23
c1a c4a c3a c5a c6a c8a c9a c10a c11a c12a c13a
hoxc_b
Electrophorus
scaffold9008
SMUG1 c4b c5b c6b c9b c11b c12b c13b
Danio
chr11
c13b SMUG1 c6b c11b c12b
hoxd_a
Electrophorus
scaffold4896
C8051512
scaffold7176
scaffold16916
d1a d3a d4a d9a d10a d11a d12a d13a
Danio
chr9
d3a d4a d9a d10a d11a d12a d13a
b
brain
spinal
kidney
heart
skeletal muscle
main EO
Sach's EO
Hunter's EO
log2 expression
0
2
4
6
8
10
12
14
c1a c3a c4a c5a c6a c8a c9a c10a c11a c12a c13a
c
E. electricus
Main
Sachs'
Hunter's
S. macrurus
E. virescens
B. brachyistius
-4 0 4
hoxc10a
hoxc11a
hoxc12a
hoxc13a

Figure 3.S1: Hox clusters and their expression patterns. (a) The arrangement of *hox* genes in *E. electricus* scaffolds relative to the *Danio* genome demonstrates both the completeness of the *E. electricus* genome sequence as well as the synteny between *E. electricus* and *D. rerio*. Gene pairs were identified based on two-way BLAST. Black arrows show the hox genes. Hollow white arrows show intervening gene models in *E. electricus* suggesting possible artifacts of the automated annotation pipeline. Hollow white boxes indicate lack of a predicted gene model but homology to the genome sequence at that location. Red box indicates a gene model which is present but which required manual adjustment of the gene model. Gray boxes indicate *D. rerio* chromosome and *E. electricus* scaffold on which the genes are located. (b) Expression of *hoxc-a* cluster in eight tissues of *E. electricus*. Normalized read counts were $\log_2$-transformed and plotted against gene location on the chromosome. (c) Subset of *hoxc-a* cluster genes overexpressed in electric organs of three Gymnotiformes and a mormyroid electric fish. Colors indicate $\log_2$ fold change between electric organ (as indicated) and skeletal muscle. Values were clamped at −4 and +4.

Figure 3.S2a

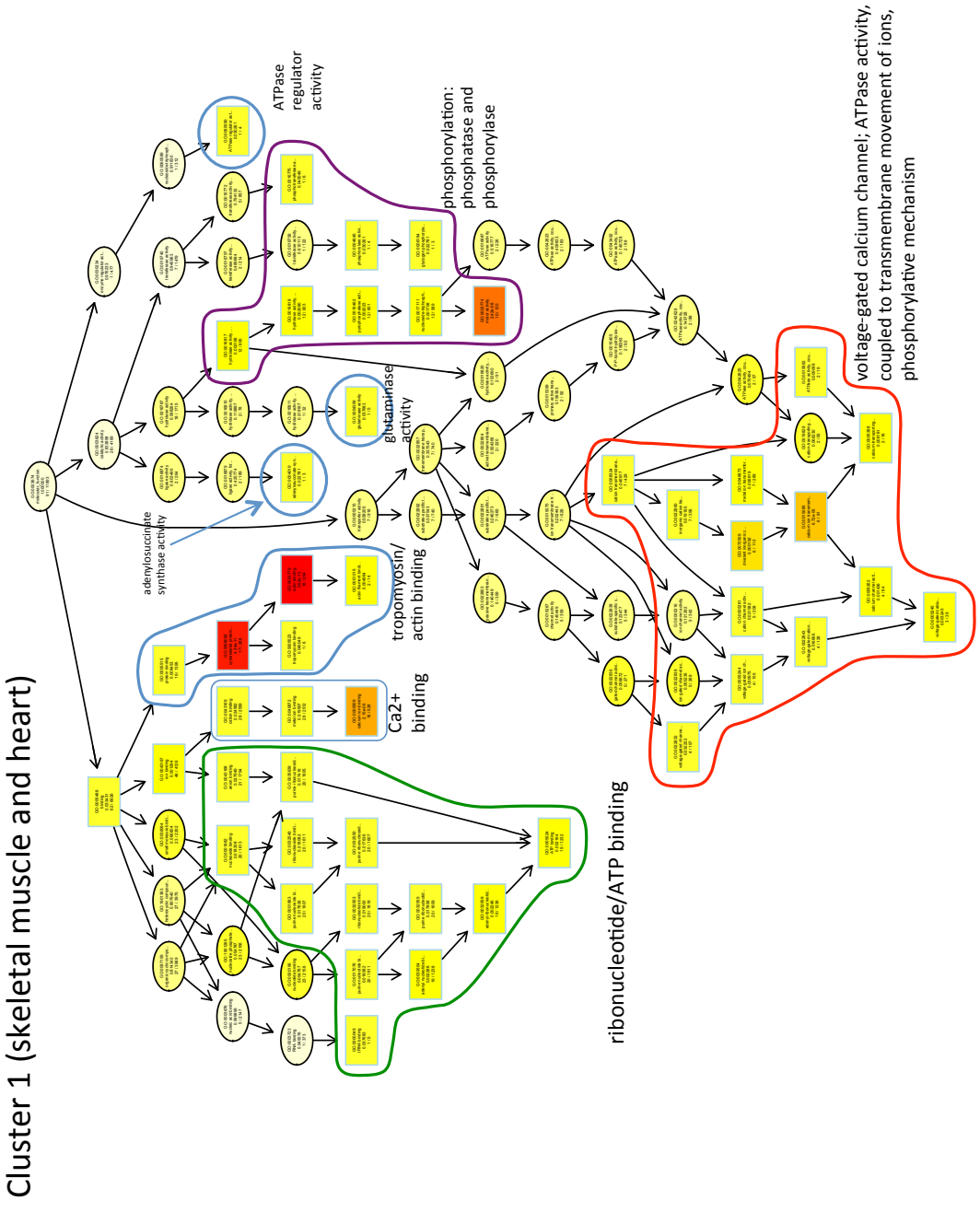

Cluster 1 (skeletal muscle and heart)
ATPase regulator activity
phosphorylation: phosphatase and phosphorylase
adenylosuccinate synthase activity
glutaminase activity
tropomyosin/ actin binding
Ca2+ binding
ribonucleotide/ATP binding
voltage-gated calcium channel; ATPase activity, coupled to transmembrane movement of ions, phosphorylative mechanism

Figure 3.S2b.

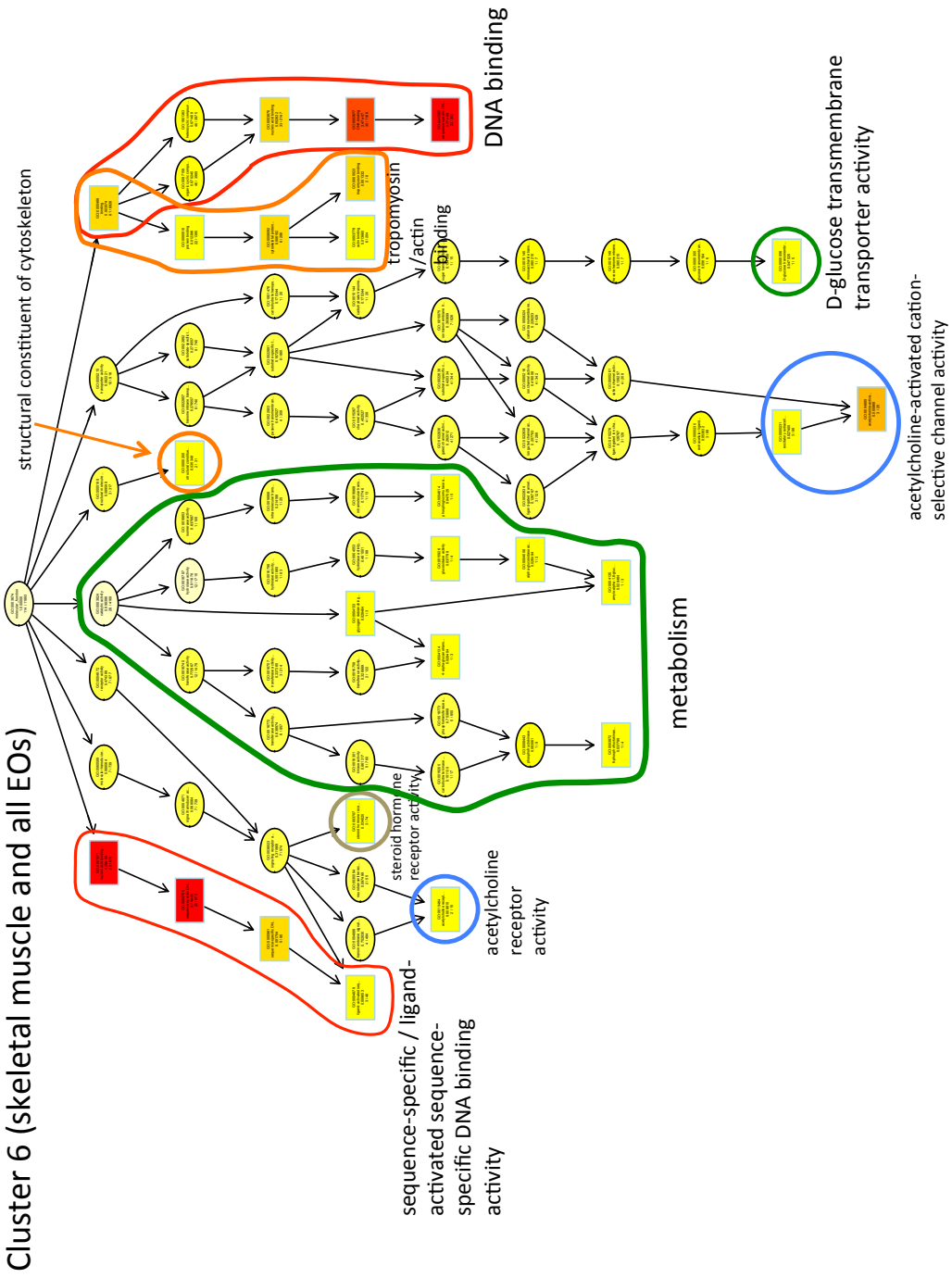

Cluster 6 (skeletal muscle and all EOs)
structural constituent of cytoskeleton
DNA binding
tropomyosin /actin binding
D-glucose transmembrane transporter activity
acetylcholine-activated cation-selective channel activity
metabolism
steroid hormone receptor activity
acetylcholine receptor activity
sequence-specific / ligand-activated sequence-specific DNA binding activity

Figure 3.S2c

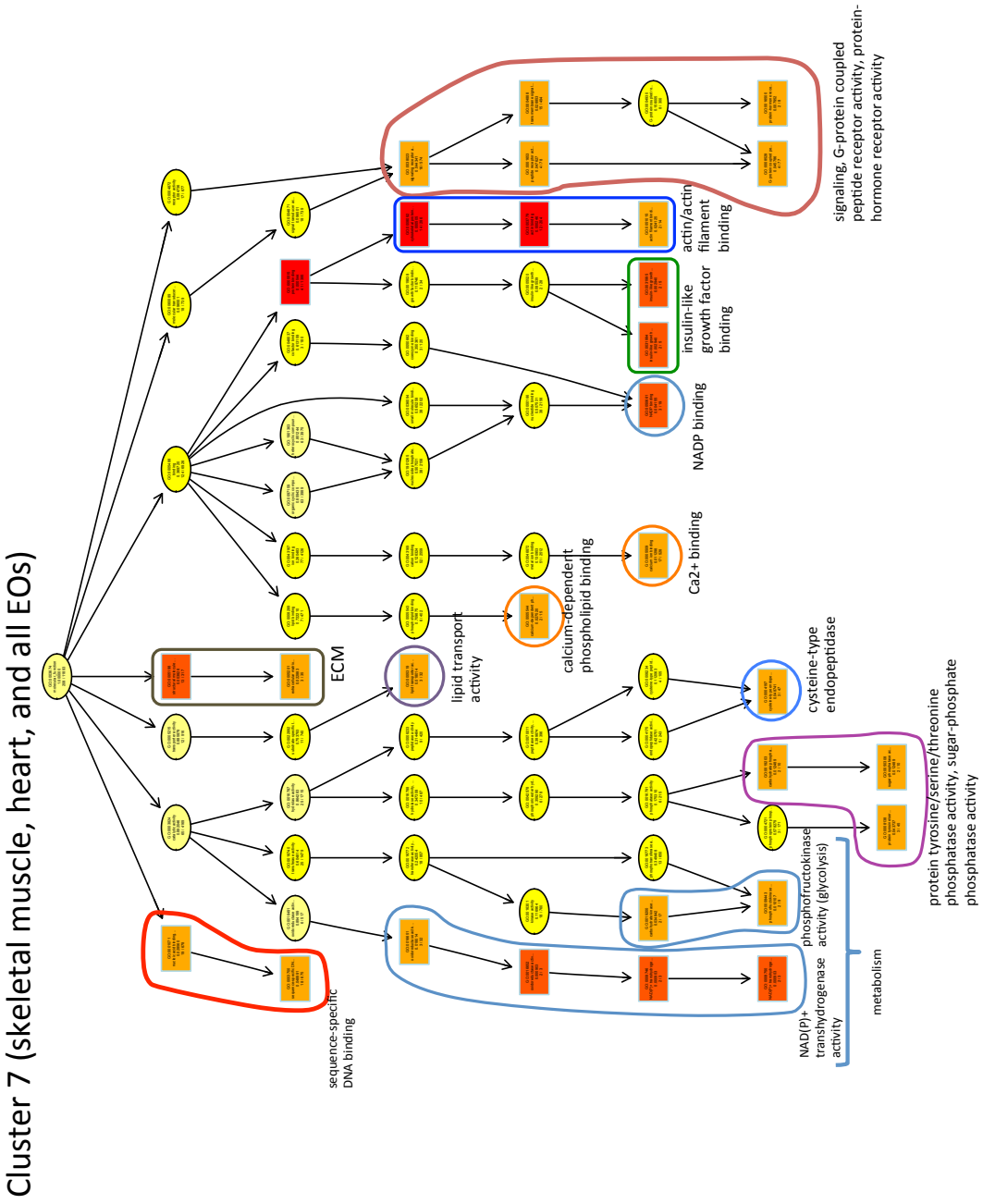

Cluster 7 (skeletal muscle, heart, and all EOs)
ECM
lipid transport activity
calcium-dependent phospholipid binding
Ca2+ binding
NADP binding
insulin-like growth factor binding
actin/actin filament binding
signaling, G-protein coupled peptide receptor activity, protein-hormone receptor activity
cysteine-type endopeptidase
phosphofructokinase activity (glycolysis)
NAD(P)+ transhydrogenase activity
metabolism
protein tyrosine/serine/threonine phosphatase activity, sugar-phosphate phosphatase activity
sequence-specific DNA binding

Figure 3.S2d

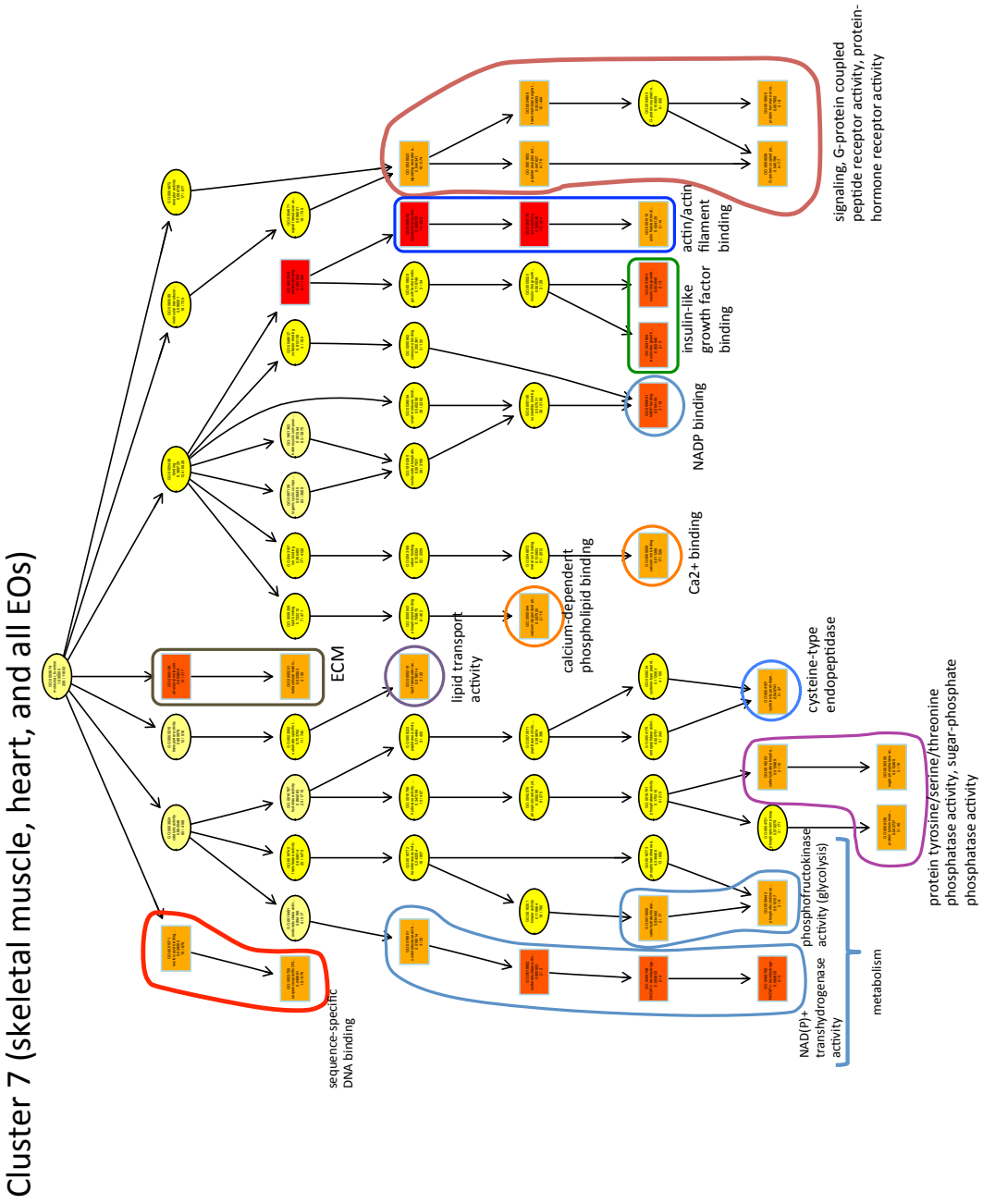

Cluster 7 (skeletal muscle, heart, and all EOs)
ECM
lipid transport activity
calcium-dependent phospholipid binding
Ca2+ binding
NADP binding
insulin-like growth factor binding
actin/actin filament binding
signaling, G-protein coupled peptide receptor activity, protein-hormone receptor activity
cysteine-type endopeptidase
protein tyrosine/serine/threonine phosphatase activity, sugar-phosphate phosphatase activity
phosphofructokinase activity (glycolysis)
NAD(P)+ transhydrogenase activity
metabolism
sequence-specific DNA binding

Figure 3.S2e

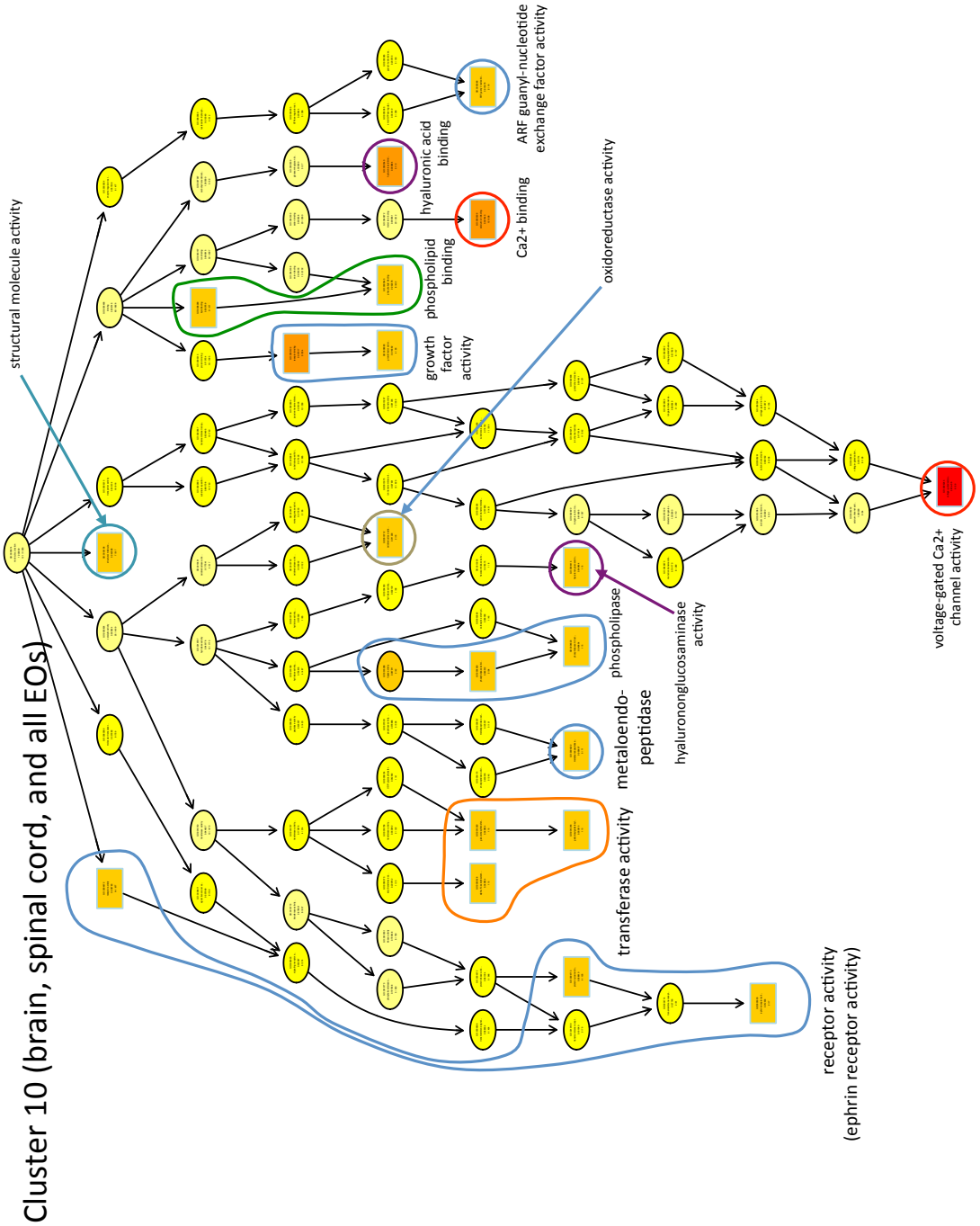

Cluster 10 (brain, spinal cord, and all EOs)
structural molecule activity
hyaluronic acid binding
ARF guanyl-nucleotide exchange factor activity
Ca2+ binding
phospholipid binding
growth factor activity
oxidoreductase activity
phospholipase
metaloendo-peptidase
hyaluronon glucosaminase activity
transferase activity
receptor activity (ephrin receptor activity)
voltage-gated Ca2+ channel activity

Figure 3.S2: Gene ontology enrichment in co-expressed genes of clusters 1, 6, 7, 9, and 10. Enriched GO terms identified using "elim" method of topGO, and a minimum node-size of 3. All GO terms present only once were removed prior to analysis. GO graphs generated from all enriched terms (p-value < 0.05 by Fisher's exact test, represented as rectangular nodes in graph) identified in each of the four EO/muscle-containing clusters. The coloration of the enriched nodes indicates p-value size, such that the smallest p-values appear dark red, and the largest p-values appear pale yellow (those closest to the p-value cutoff of 0.05). (A) Graph of GO term topology generated from the 14 enriched GO terms in cluster 1. (B) Graph of GO term topology generated from the 19 enriched GO terms in cluster 6. (C) Graph of GO term topology generated from the 20 enriched GO terms in cluster 7. (D) Graph of GO term topology generated from the 16 enriched GO terms in cluster 9. (E) Graph of GO term topology generated from the 18 enriched GO terms in cluster 10.

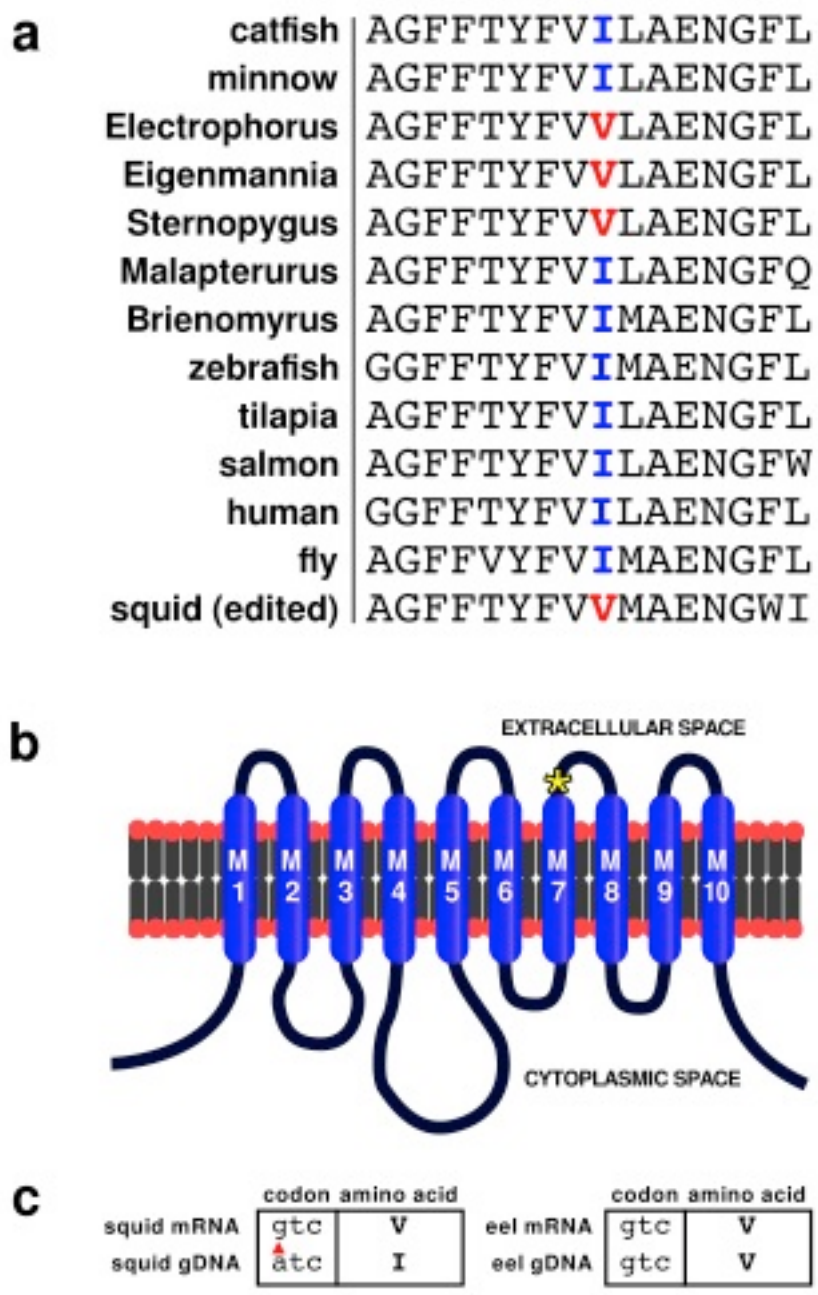


Figure 3.S3: Shared amino acid substitution in an abundant sodium pump of gymnotiform electrocytes. The α2 isoform of the sodium pump, which is highly over-expressed in the electrocyte, shows an amino acid substitution at a conserved site. (a) This substitution (red V) is present in *E. electricus*, *E. virescens* and *S. macrurus*, suggesting that

it may have occurred at the origin of Gymnotiformes. It does not occur in the mormyrid species we have studied. In an interesting case of parallel evolution, the same substitution occurs in squid. (b) Illustration indicating approximate site of amino acid substitution (yellow asterisk) within the protein topology. (c) In squid the changed amino acid is due to RNA editing rather than a permanent change in the codon. This amino acid change is thought to cause enhanced sodium transport [27].

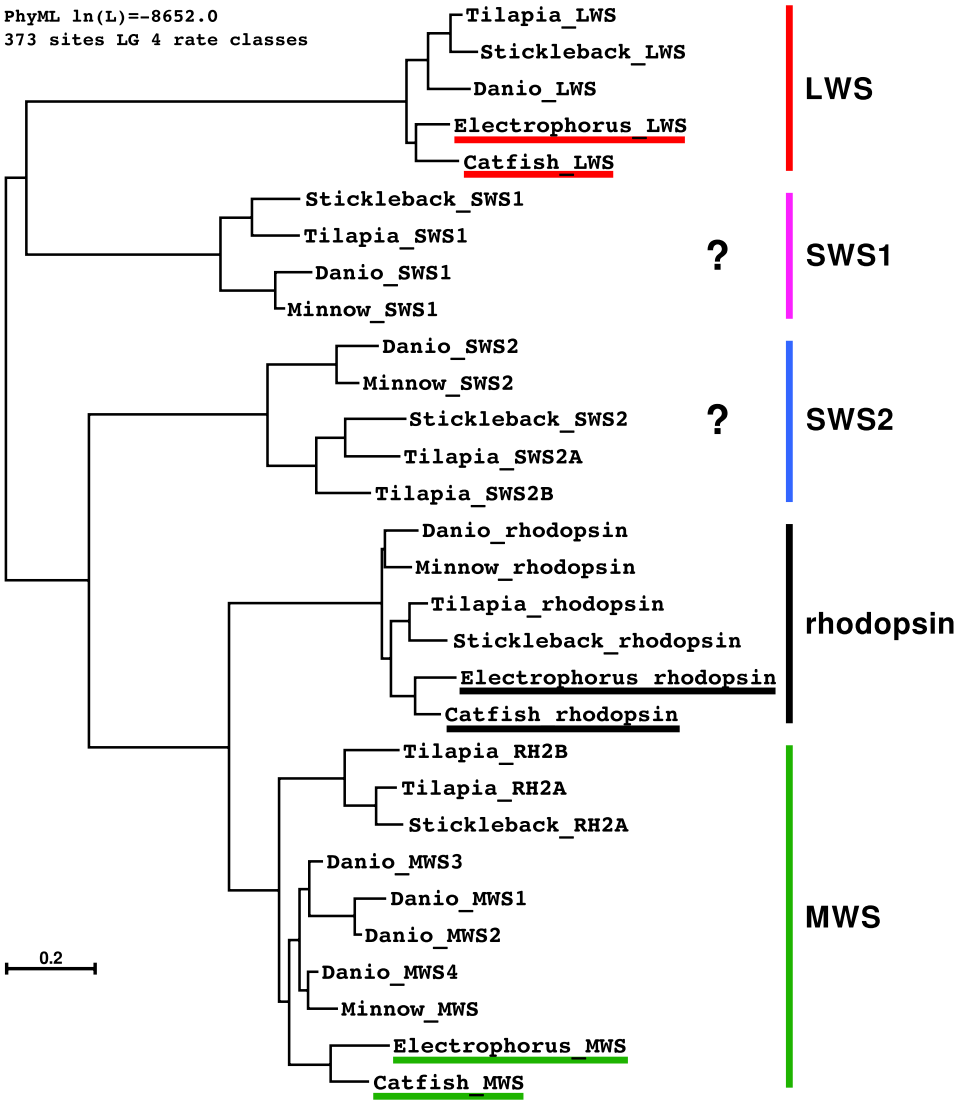


Figure 3.S4: Phylogeny of opsin genes. Phylogenetic comparisons were made between multiple related fishes for long wavelength, medium wavelength, and short wavelength opsin genes and rhodopsin gene. Both short wavelength opsins were missing from the *E. electricus* genome.

a.

E. electricus mRNA-Seq, Illumina HiSeq 2000	
Tissue	**Read count**
Brain	108433089
Spinal	114932446
Heart	116081594
Skeletal muscle	113766333
Main EO	118795559
Sachs' EO	116426272
Hunter's EO	110142469
Kidney (paired reads)	154133686

b.

E. electricus miRNA-Seq, Illumina HiSeq 2000	
Tissue	**Read Count**
Brain	13473215
Spinal	19761253
Heart	13332197
Skeletal muscle	15603748
Main EO	14274910
Sachs' EO	17304883
Hunter's EO	11974351
Kidney (paired reads)	101853001

c.

S. macrurus miRNA-Seq, Illumina HiSeq 2000	
Tissue	**Read Count**
EO	64713588
Skeletal muscle	73223556

Table 3.S1: Summary of sequencing library depths. Read counts for (a) *E. electricus* mRNA sequencing reads used in this study and in the previous study by this consortium [13] (b) *E. electricus* miRNA sequencing reads used in this study, and (c) *S. macrurus* miRNA sequencing reads used in this study.

	Electrophorus electricus	*Danio rerio*	Medaka (*Oryzias latipes*)	Platyfish (*Xiphophorus maculatus*)	Stickleback (*Gasterosteus aculeatus*)	Fugu (*Takifugu rubries*)	Green Spotted Pufferfish (Tetraodon nigroviridis)	Tilapia (*Oreochromis niloticus*)	*Homo sapiens*	Mouse (*Mus musculus*)
mean exon length	283	244	155	236	163	152	149	229	318	320
mean intron length	1099	2962	1233	1304	781	645	506	1410	7170	5493
mean CDS length	1664	1505	1454	1640	1477	1898	1638	1783	1243	1419
mean transcript length	2992	2072	1551	2583	1655	1917	1702	2534	2106	2424
mean length 5' UTR	274	170	117	302	148	111	110	363	236	205
mean length 3' UTR	1058	616	86	989	185	16	83	818	821	936
number of genes analyzed	23736/19039	26145	19686	20366	20774	18510	19589	21437	22682	22695

Table 3.S2: Comparison of gene number and structure across species. All *E. electricus* protein-coding genes were used in exon and intron length calculations (23,736 genes). Fragmented *E. electricus* gene models were removed prior to calculating coding sequence (CDS) length, transcript length, and untranslated region (UTR) lengths (19,039 genes analyzed). Gene statistics for other species were generated for protein-coding genes from whole-genome .gtf files downloaded from Ensembl (release version 70).

References

1. Albert JS, Crampton WGR. Electroreception and electrogenesis. Pp. 431–472. In: Claiborne DHEJB, Boca R, editors. The Physiology of Fishes. 3rd ed. FL: CRC Press; 2005.
2. Coates CW, Cox RT. A comparison of length and vlotage in the electric eel, *Electrophorus electricus* (Linnaeus). Zool Sci Contrib N Y Zool Soc. 1945;30:89–93.
3. Finger S, Piccolino M. The shocking history of electric fishes : from ancient epochs to the birth of modern neurophysiology. New York: Oxford University Press; 2011.
4. Bennett MVL. Electric organs. Fish Physiol. 1971;5:347–491.
5. Kirschbaum F, Schwassmann H: Ontogeny and evolution of electric organs in gymnotiform fish. J Physiol Paris 2008;102(4):347-356.
6. Moller P. Electric Fishes: History and Behavior. London: Chapman & Hall; 1995.
7. Cuellar H, Kim JA, Unguez GA. Evidence of post-transcriptional regulation in the maintenance of a partial muscle phenotype by electrogenic cells of *S. macrurus*. FASEB J. 2006;20:2540.
8. Gallant JR, Hopkins CD, Deitcher DL. Differential expression of genes and proteins between electric organ and skeletal muscle in the mormyrid electric fish Brienomyrus brachyistius. J Exp Biol. 2012;215(14):2479–94.
9. Mate SE, Brown KJ, Hoffman EP. Integrated genomics and proteomics of the Torpedo californica electric organ: concordance with the mammalian neuromuscular junction. Skelet Muscle. 2011;1(1):20.

10. Nazarian J, Berry DL, Sanjari S, Razvi M, Brown K, Hathout Y, et al. Evolution and comparative genomics of subcellular specializations: EST sequencing of Torpedo electric organ. Mar Genomics. 2011;4(1):33–40.

11. Nazarian J, Hathout Y, Vertes A, Hoffman EP. The proteome survey of an electricity-generating organ (Torpedo californica electric organ). Proteomics. 2007;7(4):617–27.

12. Patterson JM, Zakon HH. Differential expression of proteins in muscle and electric organ, a muscle derivative. J Comp Neurol. 1996;370:367–76.

13. Gallant JR, Traeger LL, Volkening JD, Moffett H, Chen PH, Novina CD, et al. Nonhuman genetics. Genomic basis for the convergent evolution of electric organs Science. 2014;344(6191):1522–5.

14. Guth R, Pinch M, Unguez GA. Mechanisms of muscle gene regulation in the electric organ of Sternopygus macrurus. J Exp Biol. 2013;216(Pt 13):2469–77.

15. Sripadi P, Nazarian J, Hathout Y, Hoffman EP, Vertes A. In vitro analysis of metabolites from the untreated tissue of Torpedo californica electric organ by mid-infrared laser ablation electrospray ionization mass spectrometry. Metabolomics. 2009;5(2):263–76.

16. Markham MR. Electrocyte physiology: 50 years later. J Exp Biol. 2013;216(Pt 13):2451–8.

17. Stoddard PK, Zakon HH, Markham MR, McAnelly L. Regulation and modulation of electric waveforms in gymnotiform electric fish. J Comp Physiol A Neuroethol Sens Neural Behav Physiol. 2006;192(6):613–24.

18. Davidson EH. Genomic Regulatory Systems: In Development and Evolution. San Diego: Academic; 2001.

19. Suemori H, Noguchi S. Hox C cluster genes are dispensable for overall body plan of mouse embryonic development. Dev Biol. 2000;220(2):333–42.

20. Pearson JC, Lemons D, McGinnis W. Modulating Hox gene functions during animal body patterning. Nat Rev Genet. 2005;6(12):893–904.

21. King BL, Gillis JA, Carlisle HR, Dahn RD. A natural deletion of the HoxC cluster in elasmobranch fishes. Science. 2011;334(6062):1517.

22. Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, et al. Zebrafish hox clusters and vertebrate genome evolution. Science. 1998;282(5394):1711–4.

23. Arnegard ME, Zwickl DJ, Lu Y, Zakon HH. Old gene duplication facilitates origin and diversification of an innovative communication system-twice. Proc Natl Acad Sci U S A. 2010;107(51):22172–7.

24. Zakon HH, Lu Y, Zwickl DJ, Hillis DM. Sodium channel genes and the evolution of diversity in communication signals of electric fishes: convergent molecular evolution. Proc Natl Acad Sci U S A. 2006;103(10):3675–80.

25. Ficker E, Taglialatela M, Wible BA, Henley CM, Brown AM. Spermine and spermidine as gating molecules for inward rectifier K+ channels. Science. 1994;266(5187):1068–72.

26. Lu Z, MacKinnon R. Electrostatic tuning of Mg2+ affinity in an inward-rectifier K+ channel. Nature. 1994;371(6494):243–6.

27. Colina C, Palavicini JP, Srikumar D, Holmgren M, Rosenthal JJC: Regulation of Na+/K+ ATPase transport velocity by RNA editing. PLoS biology 2010, 8 (11): e1000540.

28. Bowmaker JK. Evolution of vertebrate visual pigments. Vision Res. 2008;48(20):2022–41.

29. Costa MPF, Novo EMLM, Telmer KH. Spatial and temporal variability of light attenuation in large rivers of the Amazon. Hydrobiologia. 2013;702(1):171–90.

30. Veilleux CC, Cummings ME. Nocturnal light environments and species ecology: implications for nocturnal color vision in forests. J Exp Biol. 2012;215(23):4085–96.

31. Peatman E, Chao, Li.: **Auburn University**. Personal communication, received catfish EST database in 2013.

32. Ambros V. MicroRNAs and developmental timing. Curr Opin Genet Dev. 2011;21(4):511–7.

33. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, et al. MicroRNA expression profiles classify human cancers. Nature. 2005;435(7043):834–8.

34. Heimberg AM, Sempere LF, Moy VN, Donoghue PC, Peterson KJ. MicroRNAs and the advent of vertebrate morphological complexity. Proc Natl Acad Sci U S A. 2008;105(8):2946–50.

35. Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen X, et al. A uniform system for microRNA annotation. RNA. 2003;9(3):277–9.

36. Griffiths-Jones S: miRBase: microRNA sequences and annotation. Curr Protoc Bioinformatics 2010;12-9.

37. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. 2006;34(Database issue):D140–4.

38. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. 2011;39(Database issue):D152–7.

39. Kloosterman WP, Steiner FA, Berezikov E, de Bruijn E, van de Belt J, Verheul M, et al. Cloning and expression of new microRNAs from zebrafish. Nucleic Acids Res. 2006;34(9):2558–69.

40. McDaneld TG, Smith TP, Doumit ME, Miles JR, Coutinho LL, Sonstegard TS, et al. MicroRNA transcriptome profiles during swine skeletal muscle development. BMC Genomics. 2009;10:77.

41. Warren WC, Hillier LW, Marshall Graves JA, Birney E, Ponting CP, Grutzner F, et al. Genome analysis of the platypus reveals unique signatures of evolution. Nature. 2008;453(7192):175–83.

42. Goljanek-Whysall K, Sweetman D, Abu-Elmagd M, Chapnik E, Dalmay T, Hornstein E, et al. MicroRNA regulation of the paired-box transcription factor Pax3 confers robustness to developmental timing of myogenesis. Proc Natl Acad Sci U S A. 2011;108(29):11936–41.

43. Rao PK, Kumar RM, Farkhondeh M, Baskerville S, Lodish HF. Myogenic factors that regulate expression of muscle-specific microRNAs. Proc Natl Acad Sci U S A. 2006;103(23):8721–6.

44. van Rooij E, Quiat D, Johnson BA, Sutherland LB, Qi X, Richardson JA, et al. A family of microRNAs encoded by myosin genes governs myosin expression and muscle performance. Dev Cell. 2009;17(5):662–73.

45. Huang H, Xie C, Sun X, Ritchie RP, Zhang J, Chen YE. miR-10a contributes to retinoid acid-induced smooth muscle cell differentiation. J Biol Chem. 2010;285(13):9383–9.

46. Naguibneva I, Ameyar-Zazoua M, Polesskaya A, Ait-Si-Ali S, Groisman R, Souidi M, et al. The microRNA miR-181 targets the homeobox protein Hox-A11 during mammalian myoblast differentiation. Nat Cell Biol. 2006;8(3):278–84.

47. Luo W, Nie Q, Zhang X. MicroRNAs involved in skeletal muscle differentiation. Journal of genetics and genomics = Yi chuan xue bao. 2013;40(3):107–16.

48. Chiavacci E, Dolfi L, Verduci L, Meghini F, Gestri G, Evangelista AM, et al. MicroRNA 218 mediates the effects of Tbx5a over-expression on zebrafish heart development. PLoS One. 2012;7(11):e50536.

49. Sun L, Xie H, Mori MA, Alexander R, Yuan B, Hattangadi SM, et al. Mir193b-365 is essential for brown fat differentiation. Nat Cell Biol. 2011;13(8):958–65.

50. Meredith RW, Gatesy J, Emerling CA, York VM, Springer MS. Rod monochromacy and the coevolution of cetacean retinal opsins. PLoS Genet. 2013;9(4):e1003432.

51. Zhao H, Rossiter SJ, Teeling EC, Li C, Cotton JA, Zhang S. The evolution of color vision in nocturnal mammals. Proc Natl Acad Sci U S A. 2009;106(22):8980–5.

52. Northcutt RG. Taste buds: development and evolution. Brain Behav Evol. 2004;64(3):198–206.

53. Alexa A, Rahnenfuher J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics. 2006;22(13):1600–7.

54. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, et al. Rfam: updates to the RNA families database. Nucleic Acids Res. 2009;37(Database issue):D136–40.

55. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, et al. Ensembl 2012. Nucleic Acids Res. 2012;40(Database issue):D84–90.

56. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. Monatsh Chem. 1994;125(2):167–88.

57. Wyman SK, Knouf EC, Parkin RK, Fritz BR, Lin DW, Dennis LM, et al. Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity. Genome Res. 2011;21(9):1450–61.

58. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. Nat Genet. 2006;38(5):500–1.

Chapter 4. Genomic basis for the convergent evolution of electric organs

This chapter was published in the following journal article:

Gallant, J.R*., Traeger, L.L*., Volkening, J.D., Moffett, H., Chen, P-H., Novina, C.D., Phillips, G.N. Jr., Anand, R., Wells, G.B., Pinch, M., Guth, R., Unguez, G.A., Albert, J.S., Zakon, H.H.+, Samanta, M.P. +, Sussman, M.R. + Science. 2014;344(6191):1522–5.

* Desinates lead authorship

+ Designates corresponding authorship

*** All supplementary files are available in the published version of this manuscript, online.

Abstract

Little is known about the genetic basis of convergent traits that originate repeatedly over broad taxonomic scales. The myogenic electric organ has evolved six times in fishes to produce electric fields used in communication, navigation, predation, or defense. We have examined the genomic basis of the convergent anatomical and physiological origins of these organs by assembling the genome of the electric eel (*Electrophorus electricus*) and sequencing electric organ and skeletal muscle transcriptomes from three lineages that have independently evolved electric organs. Our results indicate that, despite millions of years of evolution and large differences in the morphology of electric organ cells, independent lineages have leveraged similar transcription factors and developmental and cellular pathways in the evolution of electric organs.

Background

Electric fishes use electric organs (EOs) to produce electricity for the purposes of communication; navigation; and, in extreme cases, predation and defense (1). EOs are a distinct vertebrate trait that has evolved at least six times independently (Fig. 1A). The taxonomic diversity of fishes that generate electricity is so profound that Darwin specifically cited them as an important example of convergent evolution (2). EOs benefit as a model for understanding general principles of the evolution of complex traits, as fish have evolved other specialized noncontractile muscle-derived organs (3). Furthermore, EOs provide a basis to assess whether similar mechanisms underlie the evolution of other specialized noncontractile muscle derivatives, such as the cardiac conduction system (4).

Electric organs are composed of cells called electrocytes (Fig. 1B). All electrocytes have an innervated surface enriched in cation-specific ion channels and, on the opposite surface, an invaginated plasma membrane enriched in sodium pumps, and, in some species, ion channels as well. The functional asymmetry of these cells, and their "in-series" arrangement within each organ, allows for the summation of voltages, much like batteries stacked in series in a flashlight. Although EOs originate developmentally from myogenic precursors, they are notably larger than muscle fibers (5). Further, they either lack the contractile machinery clearly evident in electron micrographs of muscle cells (Fig. 1B) or, if sarcomeres are present, as in mormyroid fish, they are disarrayed and noncontractile (Fig. 1B). Finally, electrocyte morphology varies widely: they can be long and slender, box-like, or flattened and pancake-like (Fig. 1B). Despite these differences in morphology, the three lineages of electric fish studied here share patterns of gene expression in transcription factors and pathways contributing to increased cell size, increased excitability, and decreased contractility.

Results and Discussion

We used next-generation sequencing technologies to construct a draft assembly of the Electrophorus electricus genome. Like all Gymnotiformes, E. electricus has a weak EO but is most famous for its distinct strong voltage EO. To inform gene predictions in the genome assembly, we generated short-read mRNA sequences from the main, Sachs', and Hunter's EOs, as well as the kidney, brain, spinal cord, skeletal muscle, and heart (6). This resulted in 29,363 gene models representing an estimated 22,000 protein-coding genes (table S1). Variance filtering of the gene models removed genes with low covariance among tissues, and subsequent k-means clustering (k = 12) revealed sets of tissue-specific

cotranscriptionally regulated genes (6) (fig. S1). We focused primarily on a reduced set of genes that were highly up-regulated only in EOs (cluster 9, 211 genes) or down-regulated in EOs compared with skeletal and heart muscle (cluster 1, 186 genes).

Next, we sequenced and performed de novo assembly of the transcriptomes from EOs and skeletal muscles in two other Gymnotiformes from South America (Sternopygus macrurus and Eigenmannia virescens), as well as in two other species with independently evolved EOs, a mormyroid from Africa (Brienomyrus brachyistius) and the electric catfish from Africa (Malapterurus electricus). For each species, we assigned orthology between transcripts by reciprocal BLAST searching of the set of E. electricus genes followed by manual confirmation of the matches (6). We focused on convergent properties of EOs versus skeletal muscle among lineages, and we then examined patterns of gene expression in transcription factors and developmental pathways to determine candidate mechanisms underlying these similarities (Fig. 2). We highlighted genes likely to be involved in phenotypic characteristics of electrocytes relative to muscle, including (i) down-regulation of myogenic transcriptional "profile," (ii) increased excitability, (iii) enhanced insulation, (iv) elimination of excitation-contraction coupling, and (v) large size.

We found elevated expression of several transcription factors (Fig. 2 and fig. S2) expressed early in muscle differentiation (7) that are typically down-regulated in skeletal muscle after differentiation. Six2a is of particular interest, given that it is known to target ARE promoter elements in Na+/K+ adenosine triphosphatases (8, 9). Concordant with the expression of early muscle transcription factors is the down-regulation of some transcription factors involved in muscle differentiation (e.g., myogenin and six4b) in E. electricus, B. brachyistius, and M. electricus, although not in the gymnotiform S. macrurus.

Interestingly, hey1, which is one of the most consistent highly up-regulated genes in the EOs across all groups of electric fishes, is abundant in zebrafish somites and down-regulated in mature muscle, and its overexpression in mammalian muscle precursor cells prevents their differentiation into muscle (10). Furthermore, hey1 is transiently expressed in the developing cardiac conduction pathway, and its overexpression in the heart prevents assembly of the sarcomeres (4).

A key feature of EOs is that current dissipation must be minimized and conducted unidirectionally from the EOs through the body of the fish and into the water. We noted two collagen genes, col6a6 and col141a1, that are up-regulated in EOs. The first is associated with muscle fibers, and the second is more generally expressed and ties the collagen fibers together. Collagen is deposited in the extracellular domain of basal lamina and is maintained by a cluster of molecules that span the membrane and attach to the cytoskeleton. Two of these membrane-spanning proteins, including a glycosyltransferase (gyltl1b) and dystrophin (mutations of which cause muscular dystrophy) (11), are also up-regulated in EOs and are probably involved in assembling the components that direct the flow of current.

Also, as expected, several transporters (atp1a2a or atp1a3a) and voltage-dependent ion channels (scn4aa) were highly expressed in all EOs, along with molecules that regulate them (znrf2a and fgf13a, respectively). Interestingly, the highly expressed gene encoding the α subunit of the sodium pump (atp1a2a) most closely resembles the isoform also expressed in transverse tubules (T-tubules) of muscle (12) and is abundant in the villi located within the invaginated side of the E. electricus electrocyte (13), suggesting that the uninnervated face of the electrocyte is derived from the T-tubule membrane.

A key step in the evolution of electrocytes requires disabling the excitation-contraction pathway. We noted variation in the extent to which genes for sarcomeric and sarcoplasmic reticulum–associated proteins are down-regulated in different species (Fig. 2, fig. S3, and table S2). Furthermore, mormyroid electrocytes still have sarcomere-like structures, although they appear disrupted (Fig. 1B). Despite these differences, the gene encoding the L-type calcium channel, or dihydropyridine receptor (cacna1s), which is localized in T-tubules and associated with excitation-contraction coupling in muscle, is down-regulated in all lineages. The smyds and hspb11 genes are also down-regulated in all lineages. These proteins associate with the sarcomeres, and zebrafish and mice with reduced expression or mutant gene copies have disrupted sarcomeres (14). The observed low levels of these genes in EOs suggest that they may promote disassembly of the sarcomeres, and we hypothesize that the early evolution of the EO included the down-regulation of this suite of genes, disabling contraction.

As electrocytes are much larger than muscle fibers, we hypothesized that this might be due to changes in insulin-like growth factor (IGF) signaling pathway genes (Fig. 2 and fig. S4). IGF signaling enhances body size and developmental rate in an organism-wide and tissue-specific fashion (15–18). IGF ligands are produced and released by muscle in an autocrine fashion (19), and differences in IGF signaling may result in differential growth of muscles. IGF signaling activates the insulin receptor substrate 1 protein (IRS1), which then binds to the regulatory subunit of phosphoinositide 3 kinase (PIK3) (20). PIK3 acts through distinct signaling targets to regulate cell size, cell proliferation, and protein synthesis and degradation (21). The IGF pathway is also autoregulated by a muscle-specific protein,

Fbxo40, which brings IRS1 to an E3 ligase complex. Thus, up-regulation of IRS1 is likely a key step in increasing IGF signaling activity in electrocytes.

Finally, the nuclear-envelope–related protein (Net37), abundant in cardiac and skeletal muscle tissues (22), regulates autocrine and/or paracrine release of IGF signaling and is required for myogenic differentiation of mouse myoblast cells (23). We detected electrocyte-specific up-regulation of igfII, a gene for PI3K (pik3r3b) and a net37-like gene in all lineages, as well as down-regulation of the negative inhibitor fbxo40. The net37-like protein was also recently reported to be highly expressed in the EO of another electric fish, the Torpedo ray (24). Together, the observed changes in expression in these key IGF signaling pathway genes suggest a conserved pathway among electrocytes that contributes to their increased size. The independent changes and the resulting enhancement in cell size highlight these genes as possible intracellular effectors in other insulin- or IGF-sensitive systems, as observed in male horned beetles (18).

Our analysis suggests that a common regulatory network of transcription factors and developmental pathways may have been repeatedly targeted by selection in the evolution of EOs, despite their very different morphologies. Moreover, our work illuminates convergent evolution of EOs and emphasizes key signaling steps that may be foci for the evolution of tissues and organs in other organisms.

Supplemental Online Materials

1. Animal Sources

Electrophorus electricus (*26*) - In all, three fish were used in this study, and purchased from a tropical fish dealer, Tri-County Tropicals (Richmond Hill, NY). Fish approximately 70 cm

in length were housed either individually, or with one other electric eel in aerated aquaria maintained at 26-28°C. From the first animal, the main EO was used for gDNA sequencing. From the second animal, whole brain, spinal cord, whole heart, skeletal muscle, Sachs' EO, main EO, and Hunter's EO were isolated and used for RNA sequencing. From the third animal, whole kidney was isolated and used for RNA sequencing.

Eigenmannia virescens (27, 28) - One *Eigenmannia* was purchased from a tropical fish dealer (Segrest Aquarium) and was used for this study.

Sternopygus macrurus (29) - A fresh-water species of knifefish native to South America was obtained commercially from Ornamental Fish (Miami, FL). Adult fish 30-50 cm in length were housed individually in 15- to 20-gallon aerated aquaria maintained at 25–28°C and fed three times weekly. One fish of undetermined sex was used in this study.

Malapterurus electricus (*30*) - A freshwater species of catfish, native to Africa, was commercially obtained from the Route 4 Aquarium in Elmwood Park (Elmwood Park, NJ). Adult fish were housed in groups in aerated aquaria maintained at 25-28˚C and fed daily. We dissected electric organ and muscle from two specimens.

Brienomyrus brachyistius (*31, 32*) – A freshwater mormyrid species native to central Africa and bred in the laboratory by JRG from two parents obtained commercially from Baileys Wholesale Tropical Fish (San Diego CA) was used for this study. Adult fish were housed in groups in aerated aquaria maintained at 25-28˚C and fed three times weekly. We performed RNA sequencing (one for SM and one for EO) on specimens of undetermined sex.

All procedures used followed the American Physiological Society Animal Care Guidelines, and were approved by the Institutional Animal Care and Use Committee at Cornell University, New Mexico State University, University of Texas at Austin, Michigan State University and the University of Wisconsin-Madison.

2. Genome Sequencing of Electrophorus electricus

Genomic DNA Isolation

Genomic DNA was isolated from the main EO using phenol-chloroform extraction (33) with several modifications. Tissue from the main electric organ was pulverized in liquid nitrogen using a mortar and pestle, and 1 ml solution D was added per 100 mg tissue. 500 ul aliquots of sample were loaded into 1.5 ml microcentrifuge tubes. To each of these tubes, 500 ul Tris-saturated phenol (pH 7.8), 500 ul chloroform-IAA, and 100-200 ul nuclease-free water were added. Tubes were mixed by inversion and spun in a microcentrifuge at 14,000 rpm for 5 minutes. The aqueous phase was transferred to a new tube and DNA was precipitated by addition of 1ml 100% ethanol and 100 ul 3M sodium acetate and incubation at -20°C for approximately 1 hour. DNA was pelleted using a microcentrifuge at 14,000 rpm for 10 minutes and the supernatant was discarded. DNA was resuspended in DEPC-treated water and put into a 42°C water bath for approx. 10 minutes until dissolved. 0.5 ul RNaseA (1 mg/ml) was added, and samples were incubated at room temperature for 20 minutes. Samples were subjected to a second round of phenol-chloroform extraction as described above. DNA was precipitated from the aqueous phase with EtOH/ NaAC and pelleted as described above, and pellets were dried slightly (2-4

minutes) in a sterile hood. DNA was resuspended in 50-100 ul DEPC-treated water, and tubes were combined. Sample concentration was measured on a Nanodrop spectrophotometer (Thermo Scientific) and the integrity of the DNA and absence of RNA were evaluated by running samples on a 1% agarose gel prior to submission to the University of Wisconsin-Madison Next Generation Sequencing Facility.

Library Preparation and DNA Sequencing

Genomic DNA libraries were sequenced on various platforms as summarized in Table S5.

Illumina GAIIx – Paired-end gDNA libraries were prepared using the Illumina Paired End Sample Preparation Kit following manufacturer's guidelines, with one modification: gDNA was size-selected on an Invitrogen E-gel instead of a standard agarose gel.

Illumina HiSeq 2000 – Paired-end (PE) DNA libraries were prepared using the Illumina TruSeq DNA Sample Preparation Kit following manufacturer's guidelines, with one modification: gDNA was size-selected on an Invitrogen E-gel rather than a standard agarose gel.

Illumina HiSeq 2000 – 2-5kb mate-pair libraries were prepared using the Illumina Mate Pair Library Preparation Kit v2 following manufacturer's guidelines.

3. RNA Extraction and Library Preparation for RNA-Seq in *Electrophorus electricus, Sternopygus macrurus, Eigenmannia virescens, Malapterurus electricus* and *Brienomyrus brachyistius*

Various mRNA libraries were sequenced on various platforms as summarized in Table S6.

E. electricus - Total RNA extraction was performed for each of the eight tissues following the phenol/chloroform extraction method outlined in Chomczynski and Sacchi 2006 (33),

with the following modifications. Tissue samples were ground in liquid nitrogen using a ceramic mortar and pestle. Following grinding, 1 ml per 100 mg tissue of solution D (4 M guanidinium thiocyanate, 25 mM sodium citrate, pH 7.0; 0.5%(wt/vol) N-lauroylsarcosine; 0.1 M 2-mercaptoethanol) was added, mixed, and transferred to 2 ml microcentrifuge tubes containing a single 50 mm steel bead. Tissue was homogenized on a bead mill for 2 minutes at 20 Hz at 4°C, tubes were rotated, and homogenized again for 2 minutes at 2 Hz. Tubes were centrifuged at 14,000 RPM for 10 minutes at 4°C, and supernatant transferred to fresh nuclease-free microcentrifuge tubes, 500 ul/tube, for phenol/chloroform extraction. Following the first extraction and precipitation, RNA pellets were resuspended in DEPC-treated water and DNase treated following Qiagen MinElute protocol in appendix C, with DNaseI. A second phenol/chloroform extraction was performed, followed by a second ethanol precipitation, and RNA wash step as described in Chomczynski and Sacchi (33). cDNA libraries were constructed from purified RNA using the Illumina TruSeq RNA Sample Preparation (v.2) kit (San Diego, CA). Libraries were sequenced on an Illumina HiSeq 2000 using 100bp single-end reads (1x100bp).

S. macrurus - Fish were anesthetized with 2-phenoxyethanol (1.0 mL/L) to excise ventral skeletal muscle and caudal EO under a dissecting microscope. Tissues were blotted dry, weighed, and immediately flash frozen in liquid nitrogen. Total cellular RNA was isolated from both tissues by chopping the frozen tissues into smaller pieces and subsequently pulverizing the tissues in liquid nitrogen using a mortar and pestle. Pulverized tissues were re-suspended in TRIzol reagent (Invitrogen, Carlsbad, CA) and total RNA extracted following manufacturer's instructions. To remove residual DNA, total RNA was treated with DNase I, Amplification Grade (Invitrogen) according to manufacturer's instructions. Total

RNA was then purified using phenol:chloroform:isoamyl alcohol extraction followed by isopropanol precipitation. cDNA libraries were constructed from purified RNA using the Illumina TruSeq RNA Sample Preparation (v.2) kit (San Diego, CA). Libraries were sequenced on an Illumina HiSeq 2000 using 100bp paired-end reads (2x100bp).

E. virescens – EOs were dissected by cutting the tail and removal of the skin. A piece of muscle was dissected from the back of the fish. From each sample, total RNA was extracted with RNA STAT-60 and then was treated with a procedure to remove ribosomal RNA. Libraries were sequenced on an Illumina HiSeq 2000 using 99bp paired-end reads (2x99bp).

M. electricus -Fish were euthanized by overdose of MS-222. EOs were dissected by removal of the skin and electric organ which are closely attached over the majority of the body surface. Skin and electric organ were separated using fine forceps. Trunk skeletal muscle was then dissected from directly below the sites where electric organ were dissected. This region was approximately ~2×1×0.5 cm, caudal to operculum, dorsal to lateral line. Tissue was stored in RNAlater, and total RNA was extracted using an RNeasy fibrous tissue extraction kit (Qiagen, Inc.). Samples were submitted to the Michigan State University RTSF for RNA sequencing using the TrueSeq mRNA sample preparation kit (Illumina, Inc.) Resulting libraries were sequenced on an Illumina HiSeq 2500 using 150 bp paired-end reads (2x100bp).

B. brachyistius - Fish were euthanized by overdose of MS-222. EOs were dissected by removal of the skin and muscle from the caudal peduncle, excision of the EO and spinal column, and finally removal of the spinal cord by inserting a fine pin into the vertebral column. Trunk skeletal muscle was dissected from ~2×1×0.5 cm, caudal to operculum,

dorsal to lateral line; skin removed. Tissue was immediately frozen in liquid nitrogen, then pulverized using pestle and mortar, and total RNA was extracted using TRIzol solution (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's instructions. Total RNA was PolyA+ purified using a FastTrack MAG mRNA isolation kit (Invitrogen). cDNA libraries were constructed from purified mRNA using the NEBnext Sample Kit for Illumina Sequencing (New England Biolabs). Libraries were sequenced on an Illumina HiSeq 2000 using 100 bp paired-end reads (2x100bp).

4. Draft Genome Assembly of *Electrophorus electricus*

Estimation of Genome Size

The haploid genome size of *E. electricus* has not been measured empirically, but the sizes for other species in the class Gymnotiformes are available (31, 32, 34-36). In order to obtain an empirical value of genome size based on sequence data, we ran the 'preqc' module (-preqc) of SGA assembler on error-corrected PE reads (37). The module gave an estimate of ~720 Mb by splitting the PE reads into 31-mers and analyzing their distribution. The completion of the SGA assembly (specifying the parameter OD=75, in addition to default parameters) using PE reads produced a genome of size 533 Mb after exclusion of small contigs (<2x read length). The differences in genome size estimation highlights the uncertainty of genome size predictions and methods. However, these estimated sizes for *E. electricus* genome agree well with the relationship between genome size and GC content previously determined in teleost fishes (38).

We also split the PE libraries from both strands into all possible 21-mers and evaluated their distribution. Distribution of 21-mers showed a single peak at frequency=55

(ignoring the error peak at frequency=1). This suggests that the PE reads sampled the genome at 55x k-mer coverage.

Assembly of Genomic DNA

A draft genome assembly was built from the Illumina paired end and mate pair reads using SOAPdenovo2, a de Bruijn graph-based genome assembler (39). Reads were assembled with de Bruijn graph parameter k=47, a value that produced the best N50 scaffold size. All analyses in the paper were performed using this version of the assembly. The assembled genome build is summarized in Table S7.

Transcriptome mapping-based evaluation

Because our genomic DNA was isolated from adult, differentiated tissue (main EO), we wanted to confirm the suitability of our SOAPdenovo2 genome build as a reference for E. electricus as a whole. To do this, we mapped the independent transcriptome assembly generated from RNA sequencing reads from eight E. electricus tissues (see Supplementary Materials, section 6) to the SOAPdenovo2 build using GMAP (40) We found that out of the 365,443 transcripts in our eight-tissue Trinity assembly, 357,859 (97.92%) transcripts aligned to the SOAPdenovo2 build. This high degree of overlap between the two independent assemblies is evidence that our SOAPdenovo2 genome build is suitable for subsequent genome and gene analyses.

Analysis of Assembly Quality of the SOAPdenovo2 Build

CEGMA-based evaluation: The SOAPdenovo2 genome build was analyzed using CEGMA v2.4 (41) with default parameters in order to evaluate completeness of coding regions. CEGMA located 446 (97%) of the 458 genes included in its core set. Of the subset of 248 most-conserved genes defined by CEGMA, 217 (88%) were identified as full-length and 245 (99%) were identified as either full or partial.

K-mer based evaluation: We also evaluated the SOAPdenovo2 assembly for completeness by checking what fraction of 21-mers within the PE reads were incorporated into the assembled genome, and the number of times those 21-mers were present in the assembly. Earlier we described the 21-mer distribution with a peak at frequency=55 representing the k-mer coverage of the PE reads. Therefore, it is expected that the 21-mers present with frequency ~110 in PE library are, on average, present twice in the actual genome, frequency ~165 are, on average, present thrice and so on. Hence, a comparison between 21-mer distribution of the PE reads and 21-mer distribution of assembled genome should illustrate the level of completeness of the assembled genome. Moreover, it highlights whether the missing regions are from the non-repetitive or repetitive parts of the actual genome.

The comparison between 21-mers from SOAPdenovo2 assembly and the PE reads suggests that the assembly is over 98% complete in the non-repetitive regions represented by 21-mer frequency between 40 and 80 in PE reads. The distribution tapered off below frequency of 40. The multiplicity of 21-mers in the genome increased for those with PE frequency> 100, but the rate of increase was lower than expected. This suggests that the SOAPdenovo2 assembly provides sub-optimal assembly of the repetitive and low-coverage regions.

5. Gene Annotation in Electrophorus electricus

Genome structural annotation

RNA-Seq reads from eight E. electricus mRNA tissue libraries were mapped to the genome using TopHat v2.0.4 (42) and specifying "--microexon-search -i 20 -I 50000" in addition to the default parameters. For the purpose of gene prediction, alignments from all eight tissues were merged and extrinsic hints files for "intron" and "exonpart" features were generated for input into AUGUSTUS v2.6 (43). Protein-coding gene models were predicted in all genome scaffolds at least 500 bp in length with AUGUSTUS, using human-specific parameters ("--species=human") and E. electricus extrinsic transcriptional evidence, with the following altered parameters in the otherwise default extrinsic.M.RM.E.W.cfg file:

```
Exonpart    1  .997 M   1 1e+100 RM 1 1 E 1 1e2 W 1 1.007
Exon        1     1 M   1 1e+100 RM 1 1 E 1 1e4 W 1 1
Intron      1    .3 M   1 1e+100 RM 1 1 E 1 1e6 W 1 1
UTRpart     1 1 .96 M   1 1e+100 RM 1 1 E 1 1   W 1 1
```

Additionally, "--alternatives-from-evidence=true --allow_hinted_splicesites=atac --UTR=on" was specified. Based on an initial manual inspection, CDS features predicted from AUGUSTUS were refined to use the first in-frame start codon and to choose the longest ORF except when homology to D. rerio (blastp (44) against Ensembl Zv9 build 69 (45) e-value <= 1E-20) suggested otherwise.

Functional annotation of predicted gene models

Gene symbols and functional annotations were assigned to AUGUSTUS gene models by comparison to D. rerio Zv9 build 70 protein sequences. D. rerio protein sequences were

used to search the E. electricus AUGUSTUS predicted protein sequences at an e-value cutoff of 1E-10. Results were compiled based on D. rerio gene ID and E. electricus hits were ordered based on alignment bit score. In addition, the top-scoring D. rerio genes for each E. electricus database entry were tracked and ranked. Thus, each D. rerio query was associated with an ordered ranking of E. electricus hits and each E. electricus database entry was associated with an ordered ranking of D. rerio queries. In the initial round of assignments, reciprocal best-hit pairs were identified and recorded and the E. electricus gene was assigned an annotation class of 'reciprocal'. In addition, for each D. rerio query, lower-scoring E. electricus hits for which the D. rerio gene was the top query match were checked in an iterative fashion for both adjacency to the top-scoring E. electricus gene and alignments to the D. rerio gene that overlapped by <20%. If identified, these E. electricus gene models were assigned to the same D. rerio gene and given an annotation class of 'split' indicating that they are likely to be incorrectly split from the primary gene model. This process was repeated for each locus until no additional assignments were made.

In a second round of annotation, all unannotated E. electricus genes that were positioned within 5000bp of the end of a scaffold were checked against the currently assigned genes sharing the top D. rerio query. If both fell at the end of a scaffold and D. rerio alignments overlapped by <20% they were assigned to the same D. rerio gene and give an annotation class of 'split_scaff', indicating that they likely belong to the primary gene model but are split across scaffolds. In the third round of annotation, D. rerio genes without assigned E. electricus genes were checked and if remaining unassigned matches were found that agreed by synteny with assigned E. electricus genes for immediately adjacent D. rerio genes, they were assigned to that D. rerio gene and given an annotation

class of 'synteny'. All remaining unassigned E. electricus genes which had any matches to D. rerio queries were assigned to the top D. rerio query and given an annotation class of 'secondary' to indicate the increased uncertainty in the assignment. After automated annotation, genes which were examined by hand in the course of work and had their annotations adjusted were assigned an annotation class of 'manual'.

All AUGUSTUS genes without high-scoring D. rerio matches were subsequently searched against the GenBank non-redundant protein database using blastp (44) with an e-value cutoff of 1E-10. Genes with GenBank hits were assigned to the top hit with a match class of 'nr'. Taking all together, the coding content of the genome was calculated as 22,228, or approximately 22,000, protein-coding genes with homology to published sequences (Table S1).

6. Transcriptome Assembly

For each fish, short read libraries from different tissues were combined and quality control and filtering was performed using the fastx toolkit (CSHL) as well as scythe and sickle (UC-Davis). The transcriptome assembly pipeline Trinity (*46*) (r2012-06-08) was utilized to perform de-novo transcript assembly of short reads for each fish using default parameters.

E. electricus RNA-Seq libraries were assembled in three phases with accumulation of additional experimental data.

i) All reads from three initial GA2 RNA-Seq libraries (muscle, main EO, Sachs' EO) were combined together and assembled using Trinity (default parameters).

ii) Reads from seven HiSeq tissues (brain, spinal cord, heart, muscle and three different EOs) were combined together and assembled using Trinity (default parameters). The assembly steps (i) and (ii) resulted in ~449,000 assembled transcripts.

iii) After completion of RNA-Seq experiment on kidney (conducted separately), reads from all eight *E. electricus* tissues were combined and assembled using Trinity, specifying --kmer_method meryl. This resulted in ~365,000 assembled transcripts.

E. virescens The reads from the *E. virescens* RNA-Seq experiment on EO and skeletal muscle were combined together. Execution of Trinity on the combined library was slow due to the size of the library. Therefore, we used two strategies - (i) subset of reads picked randomly from the library, (ii) digital normalization (http://ged.msu.edu/angus/diginorm-2012/tutorial.html) - to reduce the number of reads, and then performed Trinity assembly of the reduced library using default parameters. All downstream evaluations were performed based on the combined set of contigs generated from these assemblies.

S. macrurus - The *S. macrurus* transcriptome assembly contained 326,623 sequence contigs representing 221,914 subcomponents. Of these sequences, 163,477 were at least 500 bp in length and 63,408 were at least 2,000 bp in length. The average sequence length was 1287 bp.

M. electricus - Trinity assemblies were generated from SM/EO paired end reads using default parameters with the Trinity 2013 Nov 10th release. The assembly contained 181,633 transcript contigs.

B. brachyistius - Trinity assemblies were generated from SM/EO paired end reads using default parameters with the Trinity 2012 October 5th release. The assembly contained 147,923 transcript contigs.

7. Calculation of Expression Levels

E. electricus - Tissue-specific read counts were generated for each AUGUSTUS gene model using the previously described TopHat alignments and the htseq-count command from HTSeq ("-m intersection-strict -a 3 -t exon -s no -i gene_id") (*47*). Reads were normalized for library size using DESeq v1.10.1 with default options (*48*). To facilitate examination of individual gene expression, library-normalized read counts from above were additionally normalized by transcript size to give "reads per kb transcript" (Table S1).

S. macrurus – Short reads from individual skeletal muscle and EO libraries were individually mapped to the Trinity transcriptome assembly using Bowtie (*49*) (v.0.12.8) with default parameters and read counting and ambiguity resolution were performed using RSEM (*50*) (v.1.2.3). Read counts were subsequently normalized for library size using the geometric mean method from DESeq (Table S1).

E. virescens, B. brachyistius – Expression data was calculated as for *S. macrurus* above (Table S1).

M. electricus – Short reads from skeletal muscle and EO libraries were independently mapped to the Trinity transcriptome assembly with RSEM (*50*), using the rsem-calculate-expression command, specifying paired-end data in addition to default parameters. All transcripts with mean read counts of < 10 across both EO and muscle were removed from analysis. Subsequently, read counts were normalized for library size using the geometric mean method from DEseq (Table S1) (*48*).

Sensitivity of E. electricus expression values

Rarefaction analysis: We performed a rarefaction analysis to evaluate the rate at which we could expect additional sequencing to result in detection of additional transcripts. To do

this, we mapped all RNA-Seq reads from main EO in *E. electricus* to the SOAPdenovo2 genome assembly using TopHat (*42*) as previously described, and from here, sampled mapped reads representing 2%, 4%, 6%, 8%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100% of the total number of mapped reads in main EO, and generated read counts for all AUGUSTUS gene models using the htseq-count command from HTseq (*47*) as previously described. We looked at what number of genes met an arbitrary threshold of 50 reads, 100 reads, and 1000 reads mapped to them in each of our 14 bins, and graphed our results (Figure S5). Our results indicated that the rate of new genes detected slows as a function of number of reads used, and our graph indicates that few new genes are detected with each incremental increase in read numbers as we approach 100% of reads used. Based on this, we feel confident that we have achieved sufficient sensitivity given available resources.

Incremental changes to read counts: In order to determine the effect of incremental changes in numbers of reads used on the stability of our gene expression values, we mapped all RNA-Seq reads from each of the *E. electricus* tissues independently to the SOAPdenovo2 genome assembly using TopHat (*42*) as previously described. Bins containing mapped reads were created by randomly sampling mapped read equally from each of the eight tissues, to reach final read counts of 2, 4, 6, 8, 10, 20, 30, 40, 60, 70, 80, 90, and 100 million reads. Read counts for each of our AUGUSTUS gene models were generated using the htseq-count command from HTseq as previously described, and we generated expression values in FPKM for each of our bins. We looked at what percentage of genes fell within 5%, 10%, 15%, 20%, 40%, 60%, and 80% of the expression values for the next highest bin (for example, expression values at 40 million reads were compared to that of 50

million; the 100 million read bin was compared to a second independently-sampled 100 million read bin). In order to reduce some of the noise from genes that were either not expressed, or very lowly expressed in our analysis, we removed first genes that had FPKM expression values < 0.05 in the two 100 million read bins (26,227 genes considered), and second, genes that had FPKM expression values of < 0.05 in any of the read bins (22,276 genes considered) (Figure S6). Our resulting graphs indicate that incremental changes in read depth have decreasing effects on the stability of expression values as you approach 100 million reads. In other words, incremental changes do not have a large effect on our gene expression values, indicating that our expression values are relatively stable.

8. K-mean Clustering of E. electricus Gene Expression

Genes with an average normalized read count across tissues of less than 10 were removed from further consideration. A variance stabilizing transformation was applied to the remaining genes using DESeq, followed by median centering. Low-information genes were removed using a local R implementation of the SUMCOV variance filter as described previously (*51*). The remaining approximately 6,000 genes were subjected to k-means clustering in R (*52*) using the stats package to discover tissue-specific expression signatures (Fig. S1). The value of k=12 was chosen after testing clustergrams, as well as iterative k-means clustering with a range of k values, to find a parameter which produced informative and stable clusters.

9. Assignment of Orthologs Between Electric Fish

S. macrurus - The one-to-one gene assignment table between *E. electricus* and *D. rerio* was used to identify orthologous genes from *S. macrurus*. *S. macrurus* transcripts assembled from Trinity were compared separately against all genes from *E. electricus* and *D. rerio*

using BLAST (*44*) (blastn, e-value: 1E-5). Subsequently, all *S. macrurus* matches with each *E. electricus* gene were ranked from highest to lowest BLAST score. Another similar table was created for each *D. rerio* gene. If the same *S. macrurus* transcript matched a pair of genes from the *E. electricus* and *D. rerio* one-to-one assignment table, it was selected as the *S. macrurus* ortholog for the pair. If different transcripts matched a pair, they were both aligned with both *E. electricus* and *D. rerio* genes using CLUSTAL and the one with highest percentage match among four alignments was selected as the *S. macrurus* ortholog. Further manual confirmation of the matches was performed for the set of genes discussed in the manuscript by aligning translated amino acid or nucleotide sequences from the electric fishes and various other species (primarily zebrafish: *D. rerio*; channel catfish: *Ictalurus punctatus*; tilapia: *Oreochromis niloticus*; stickleback: *Gasterosteus aculeatus*; western clawed frog: *Xenopus tropicalis*; human: *Homo sapiens*), and generated gene trees using a maximum likelihood criterion using SeaView (*53*).

<u>*E. virescens*</u> - An identical procedure as *S. macrurus* was used to assign orthologous genes in *E. virescens*.

<u>*M. electricus*</u>- An identical procedure as *S. macrurus* was used to assign orthologous genes in *M. electricus*.

<u>*B. brachyistius*</u> - The *B. brachyistius* SM/EO paired-end assembly contained 147,923 assembled transcript contigs. Of these, 62,417 had homologues in the NCBI nr database using blastx and a cutoff of 1E-10. Of the 147,923 sequences for *B. brachyistius*, 49,902 had a homologous sequence in *E. electricus* (using tblastx, cutoff 1E-10). About 69% of the reduced set of *E. electricus* transcripts had a match in *B. brachyistius* sequences. Reciprocal

blast between the reduced *E. electricus* transcriptome and the *B. brachyistius* transcriptome yields 7,960 matches between the two.

10. Analysis of Muscle Protein Regulation in Electric Organs

Data analysis approach

The transcripts of more than 30 muscle genes that are directly involved in muscle contraction, i.e., sarcomere, T-tubule, and sarcoplasmic reticulum (SR) (*54-56*), were identified in the transcriptomes of *E. electricus*, *S. macrurus*, *E. virescens*, *B. brachyistius,* and *M. electricus*. Gene identities for each transcript were derived using SeaView (*53*) (v.4.4.0) with ClustalW (*57*) alignments of the protein-coding regions of the electric fish transcripts and teleost orthologs retrieved from the Ensembl database (*44*). The expression level for each gene was analyzed, and the abundance of each transcript in EO relative to that in skeletal muscle (SM) was computed for each species.

Results

Expression of most contraction-related genes was detected in all five species of electric fish and is shown in Table S2 and Figure S3. Transcripts with very low expression after normalization or those not detected included nebulette, JSRP1 (junctional sarcoplasmic reticulum protein 1), and HRC (histidine rich calcium binding protein), and are not shown. The relative transcript profiles of contraction-related genes were similar in all three EOs of *E. electricus*, i.e., main, Hunter's and Sachs' and showed that >65% of these genes were down-regulated at least 4-fold in the EOs compared to skeletal muscle. The transcript profiles found in *B. brachyistius* and *M. electricus* were similar to that of *E. electricus* in that most (~66% for both species) contraction-related genes were down-regulated in the EO by at least 4-fold. In contrast to both *E. electricus*, *B. brachyistius,* and *M. electricus*, the EO

and skeletal muscle of *S. macrurus* showed similar expression levels in ~90% of contraction-related genes analyzed, whereas only one gene (troponin-C1) was found to be down-regulated by at least 4-fold in EO. The EO of *E. virescens* showed downregulation of contraction-related genes to a lesser extent than that found in *E. electricus*, *B. brachyistius,* and *M. electricus* but greater than in *S. macrurus*. Specifically, ~56% of these were down-regulated by at least 4-fold compared to skeletal muscle whereas ~42% showed similar expression. In the EOs of all five species less than 5% of contraction-related genes were found to be up-regulated by 4-fold or more compared to skeletal muscle. The transcript profiles of contraction-related genes found in the EOs of these five species are consistent with the expression profiles of the myogenic regulatory factor (MRF) genes of the MyoD family (Fig. S2), which regulate the transcription of many of these contraction-related genes in mammalian skeletal muscle since MRFs are strongly down-regulated in EOs of *E. electricus*, *B. brachyistius*, and *M. electricus* less so in the EO of *E. virescens*, and not down-regulated in the EO of *S. macrurus* (Fig. S2). These data indicate a transcriptional program that is more similar between the EO and skeletal muscle of *S. macrurus* than that observed in *E. electricus*, *B. brachyistius,* or *M. electricus*. These observations are intriguing in view of the evolutionary relationships between these electric fish species. Despite sharing a common gymnotiform ancestry, the contractile muscle gene expression profiles in EOs of *E. electricus* and *S. macrurus* differ considerably, whereas the EO of the distantly related *B. brachyistius* and *M. electricus* exhibits both contraction-related and MRF gene expression patterns that are similar to those in *E. electricus* but not those in *S. macrurus*. Further, these contraction-related gene expression patterns cannot be used to predict the cellular ultrastructure of electrocytes in these electric fish species. Previous studies have shown

that fully differentiated electrocytes of *B. brachyistius* contain myofilamentous structures (*58, 59*) whereas mature electrocytes of *S. macrurus* do not (*5*, *60*, *61*). Myofilamentous structures are observed in electrocytes of young but not adult *E. electricus* (*62, 63*). In sum, these data indicate that transcriptional repression of the myogenic program is not a requisite for the emergence and maintenance of the EO phenotype.

11. Author Contributions

GAU, HZ, JRG, and JSA provided organisms; GAU, HZ, JRG, and LLT isolated the RNA and DNA; JRG and LLT designed and supervised the preparation involved in sequencing RNA and DNA; JDV, JRG, LLT, GNP, and MPS conceived and executed the data analysis pipeline; CDN, GAU, GBW, HM, HZ, JDV, JRG, LLT, MP, MPS, P-H C, RA, and RG analyzed the data; CDN, HM, JRG, LLT, MPS, and MRS designed the experiments; JRG, LLT, HZ and MRS wrote the manuscript and all authors edited the manuscript; MPS and MRS supervised this project.

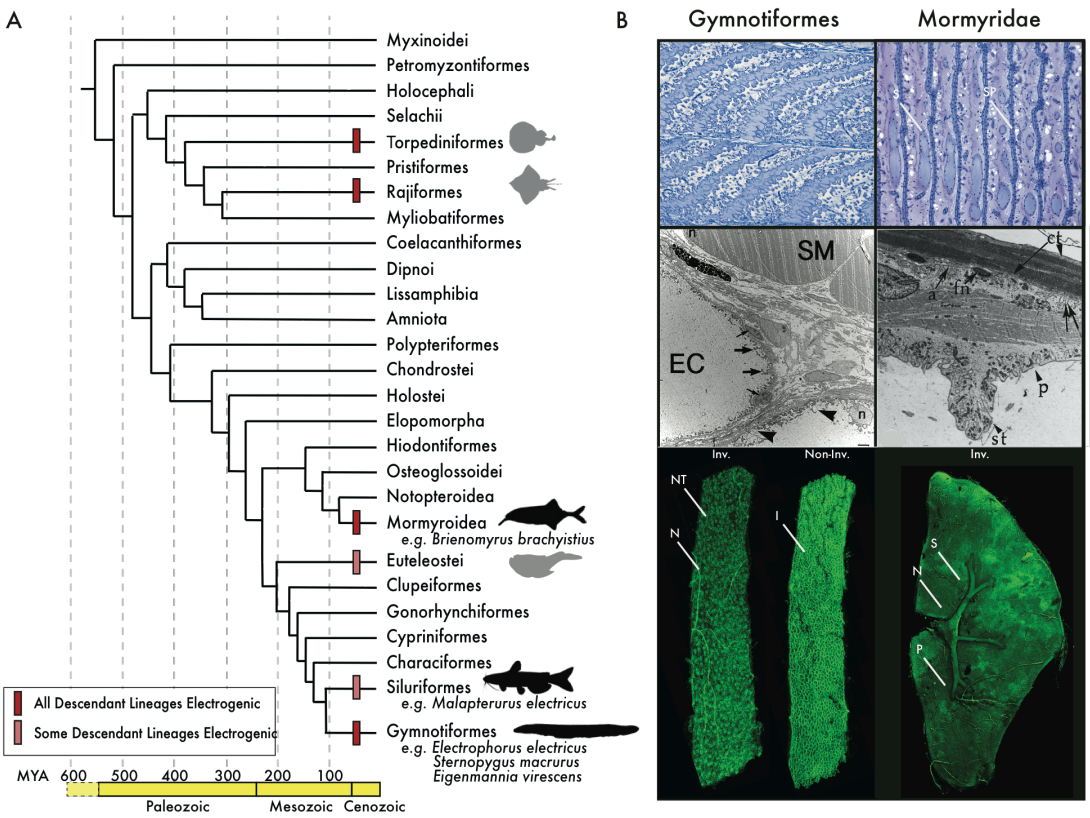


Fig. 4.1. Origins and diversity of EOs in vertebrates.

(**A**) Phylogenetic tree of vertebrate orders and major groups of electric fishes, after (*25*). Geological periods and ages [in million years ago (MYA)] are shown at bottom. The origins of electrogenesis are indicated with bars (see legend) at internal branches. Black silhouettes denote lineages surveyed in the present study; gray silhouettes represent electrogenic lineages that were not surveyed. (**B**) (Top left) Sagittal sections through the *E. electricus* EO for the innervated, invaginated face and uninnervated smooth faces of the electrocyte and their in-series arrangement. (Top right) Sagittal section through the EO of the mormyroid *Paramormyrops kingsleyae.* Anterior is left; posterior is right. In mormyroids, innervation is restricted to a narrow region of the stalk system (S) protruding from the innervated, anterior face of the electrocyte. Also note the central filament of

sarcomeric proteins (SP) between the multinucleated electrocyte faces. (Middle left) An electron micrograph of both skeletal muscle (SM) and electrocytes (EC) from the gymnotiform *S. macrurus*, which contain an amorphous cytoplasm devoid of sarcomeres: the striated, contractile structures that fill the cytosol of muscle cells. Peripheral nuclei (n) are marked in both electrocyte and muscle cells. In electrocytes, thick arrows point to mitochondria, thin arrows point to satellite cells, and arrowheads mark membrane-bound vesicular structures. Scale bar, 2 μm. (Middle right) An electron micrograph of an electrocyte of the mormyroid *P. kingsleyae*, illustrating the disorganized sarcomeric proteins in the center of the electrocyte. The outer edge of an electrocyte forms a "footplate" that apposes the connective tissue sheath (ct) surrounding the EO. The anterior face (a) of the electrocyte forms the major surface of the plate lying against the connective tissue surface. Fibroblast nuclei (fn), papillae (p), and stalk (st) are also indicated. Double arrows correspond to invaginations of the posterior face. Scale bar, 4 μm. [Image provided by Andrew Bass (Cornell University)] (Bottom left) A confocal reconstruction of an *E. electricus* electrocyte from anterior and posterior views. The nerve (N) innervating the innervated (Inv.) face is clearly visible, along with the many cholinergic nerve terminals (NT). The numerous invaginations (I) of the noninnervated (Non-Inv.) face are visible. (Bottom right) A confocal reconstruction of a *P. kingsleyae* electrocyte, clearly showing the protruding stalk system (S) from the anterior face. The stalk junction is innervated by motoneurons (N) in a highly localized fashion to contrast with *E. electricus*. Penetrations (P) are also visible in the electrocyte face.

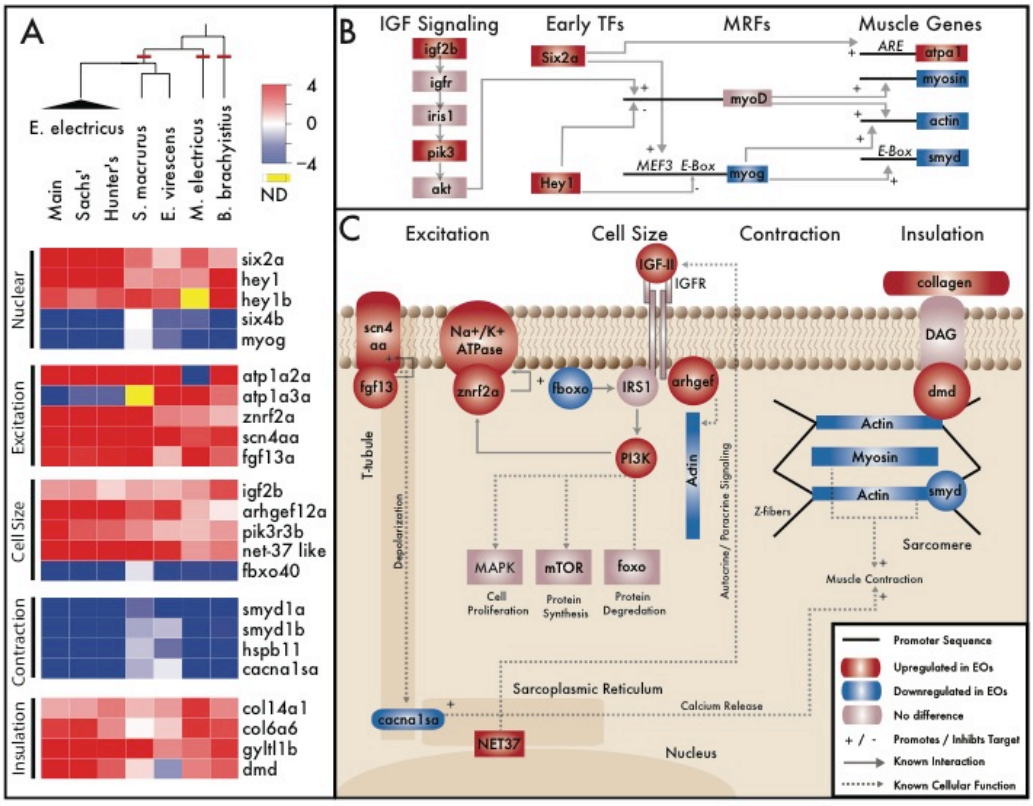


Fig. 4.2. Common toolkit for convergent evolution of EOs. (**A**) RNA-Seq was performed on five species, representing three independent origins of electrogenesis (cladogram, red lines). Also shown are plots of the log-transformed ratio of EO to skeletal muscle expression genes (red, up-regulated in EO; blue, down-regulated in EO) in several categories of function, including (i) nuclear transcription factors, (ii) genes that regulate cell excitation, (iii) genes that regulate cell size, (iv) genes involved in contraction and excitation contraction coupling, and (v) genes encoding proteins that surround individual electrocytes to provide the scaffold for insulation. *hey1b* data for *E. electricus* was derived from the Trinity transcriptome assembly (6). (**B**) Interaction of identified IGF signaling and transcription factors (TFs). IGF signaling pathway genes and early TFs influence the expression of muscle regulatory factors (MRFs), which ultimately lead to the expression of

muscle-specific effector genes (table S3). (**C**) Interactions of genes identified in (A) are shown, grouped by function. For each, we list known patterns of expression in electric fish or the result of knockout studies in other vertebrates (table S4). IGFR, IGF receptor; MAPK, mitogen-activated protein kinase.

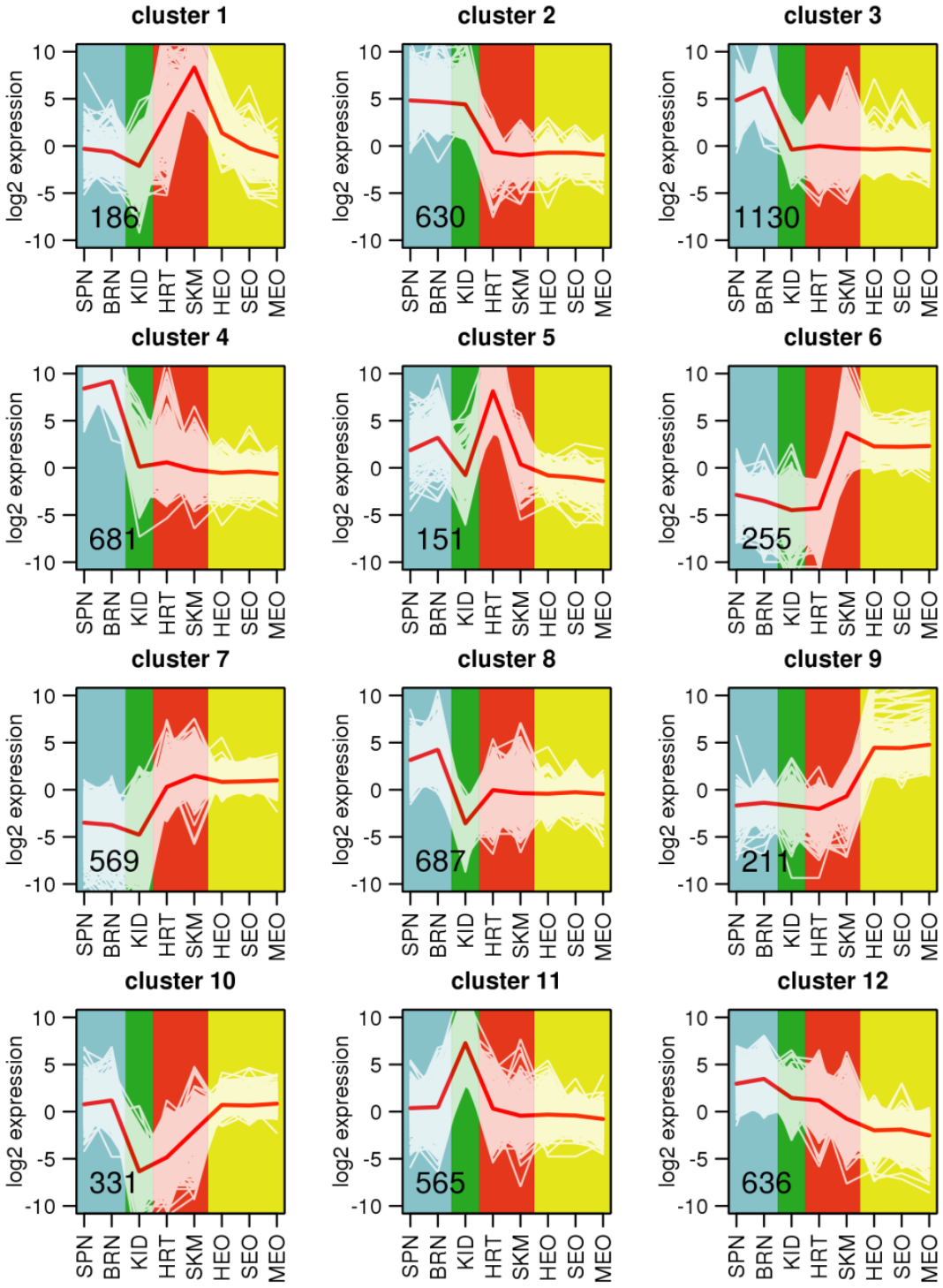


Fig. 4.S1. Clustering of co-expressed genes in *E. electricus*. A k-means clustering analysis (k=12) was performed as described in the Supplementary Text. Values in lower-

left indicate the number of genes in each cluster. White plot lines represent individual genes and red plot lines show median values for the cluster. Background shading indicates general categories of tissue/cell type. SPN=spinal cord; BRN=brain; KID=kidney; HRT=heart; SKM=skeletal muscle; HEO=Hunter's EO; SEO=Sachs' EO; MEO=main EO.

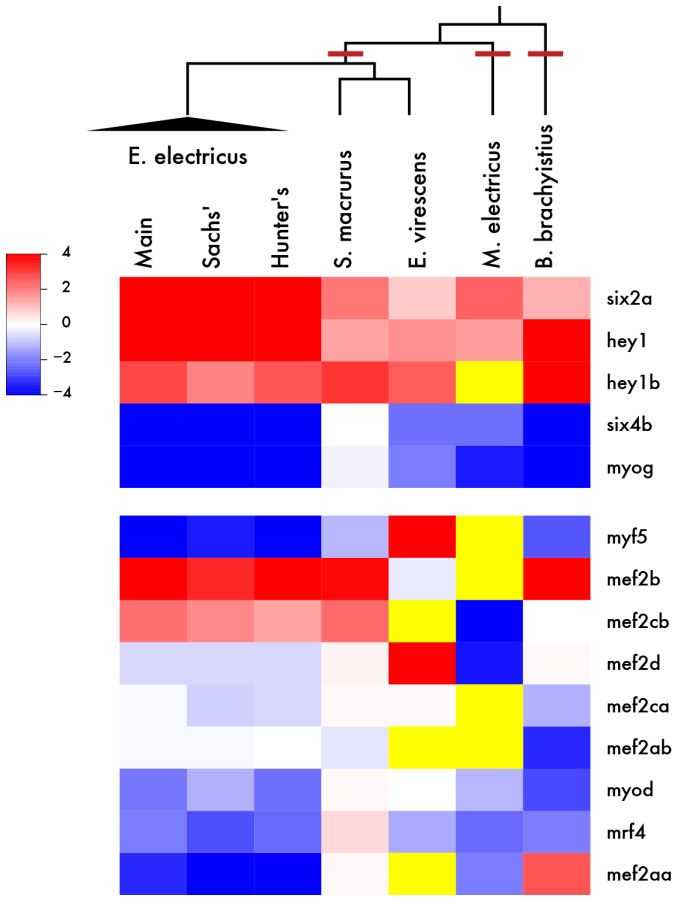


Figure. 4.S2 Heatmap of myogenic transcription factor and related muscle development gene expression. Log2 transformed ratios of electric organ to skeletal muscle expression are displayed. Red colors indicate genes highly upregulated in electric organ, blue colors indicate genes highly downregulated in electric organ for each species. Yellow colors indicate the transcript was not detected. Note that myogenin is the sole "classical" transcription factor that is consistently down-regulated in all species.

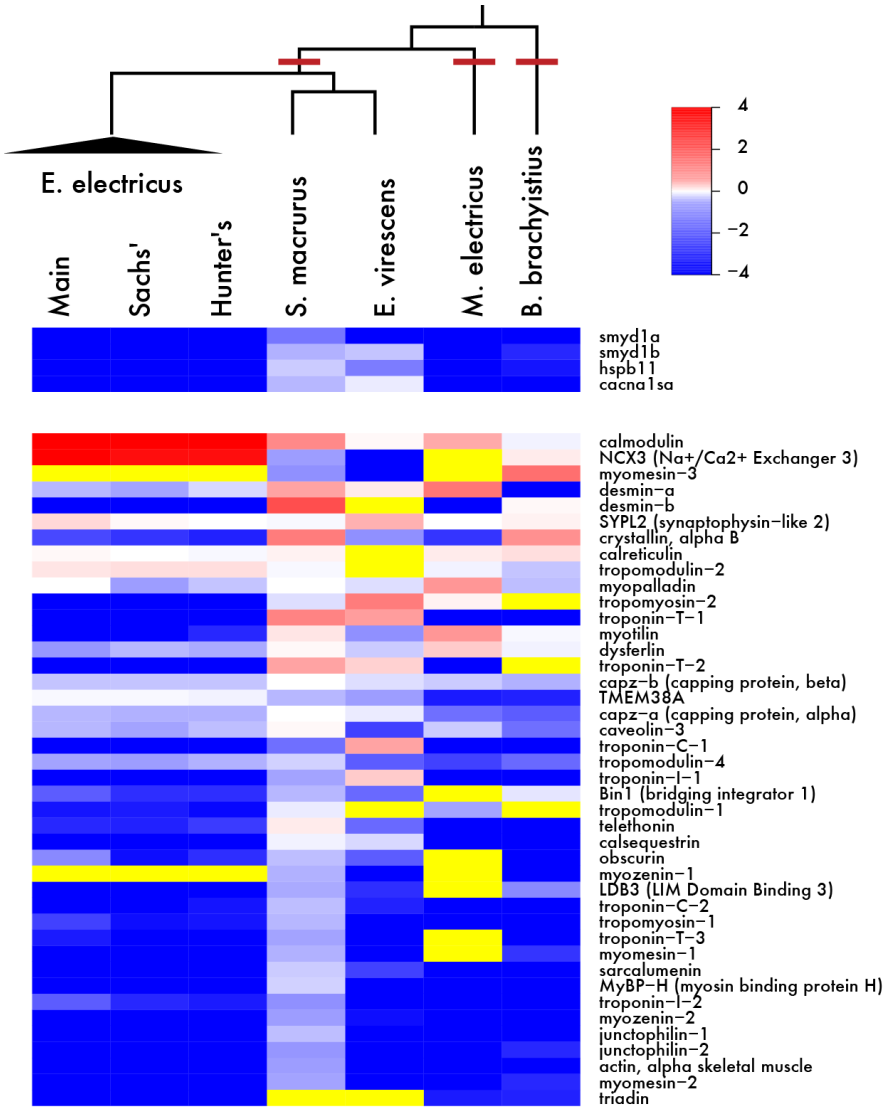


Fig. 4.S3 Heatmap of sarcomeric gene expression. Log2 transformed ratios of electric organ to skeletal muscle expression are displayed. Red colors indicate genes highly upregulated in electric organ, blue colors indicate genes highly downregulated in electric organ for each species. Yellow indicates the transcript was not detected.

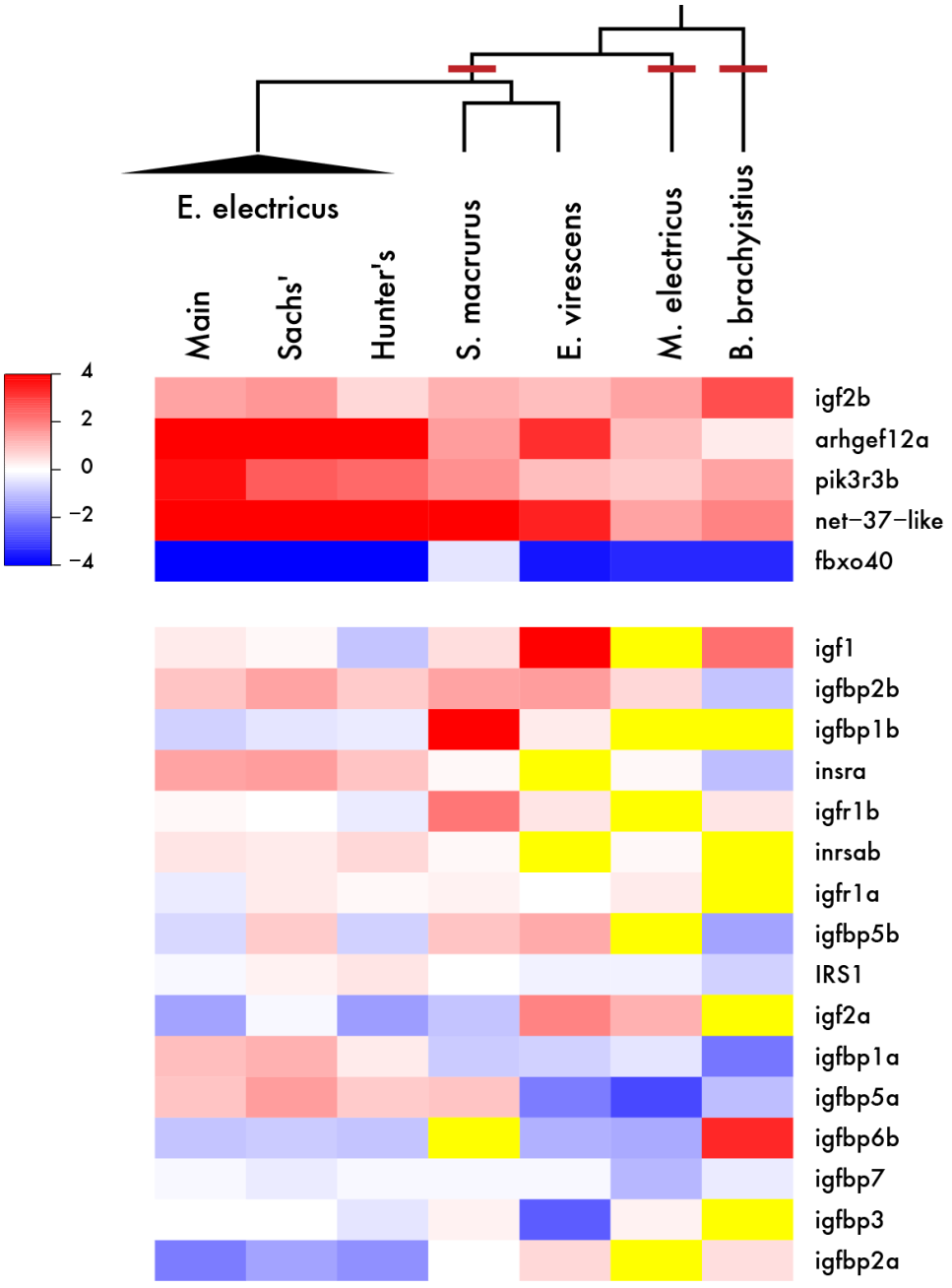


Fig. 4.S4 Heatmap of IGF signaling gene expression. Log2 transformed ratios of electric organ to skeletal muscle expression are displayed. Red colors indicate genes highly upregulated in electric organ, blue colors indicate genes highly downregulated in electric organ for each species, yellow indicates the transcript was not detected. The set of genes

that are similarly regulated in all species and that are illustrated in Fig. 2 are igf2b, arhgef21a, pik3r3b, NET37-like, and fbxo40.

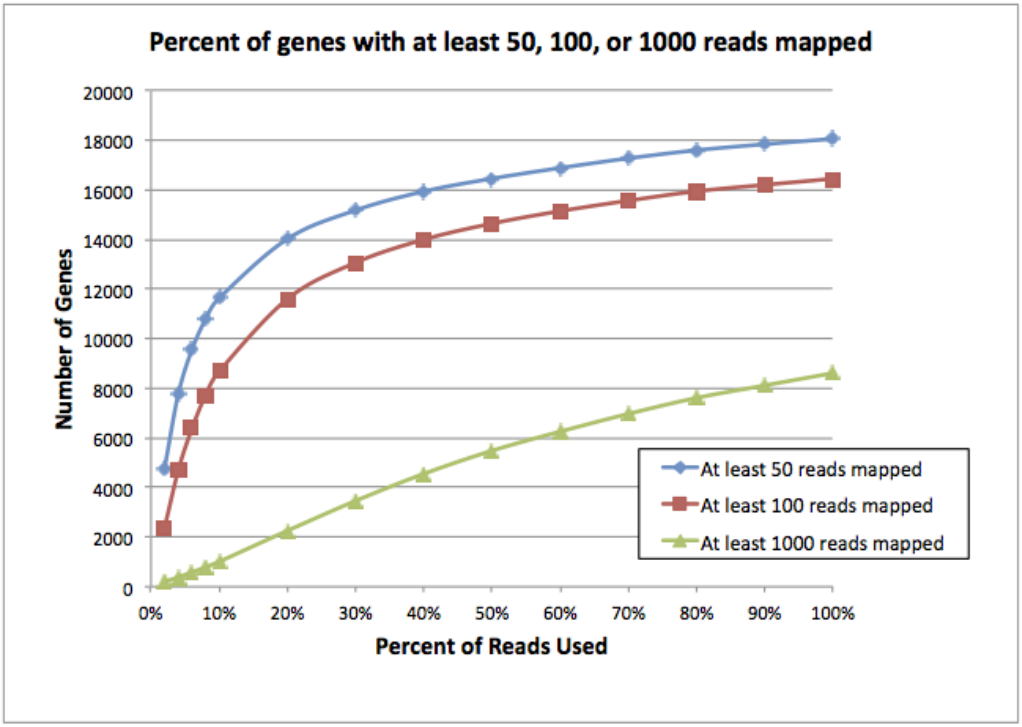
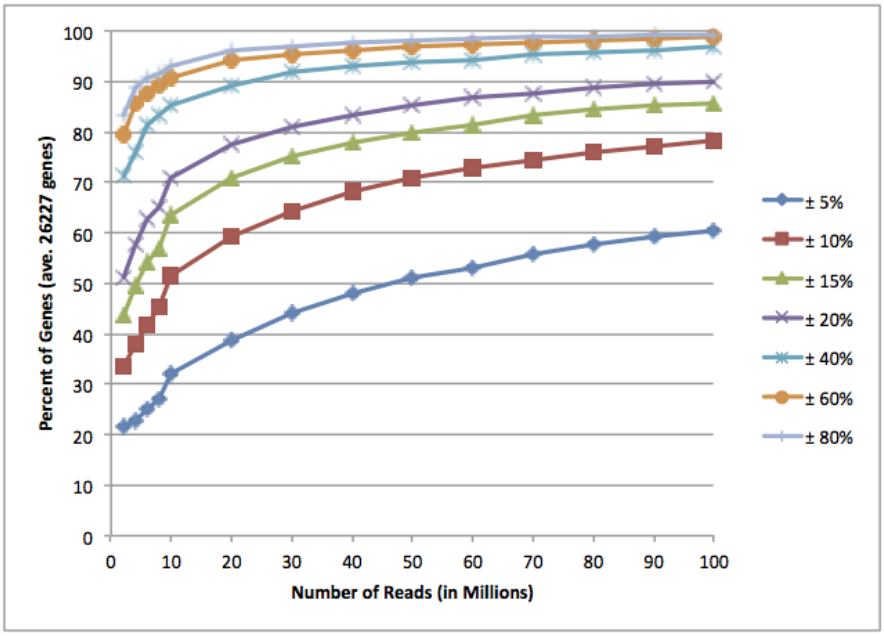


Fig. 4.S5. Rarefaction analysis of additional sequencing depth on transcript content. We mapped all RNA-Seq reads from main EO in eel to the SOAPdenovo2 genome assembly using TopHat (*42*) as previously described, and from here, sampled mapped reads representing 2%, 4%, 6%, 8%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100% of the total number of mapped reads in main EO, and generated read counts for all AUGUSTUS gene models using the htseq-count command from HTseq (*43*) as previously described. We looked at what number of genes met an arbitrary threshold of 50 reads, 100 reads, and 1000 reads mapped to them in each of our 14 bins.

4.S6a



4.S6b

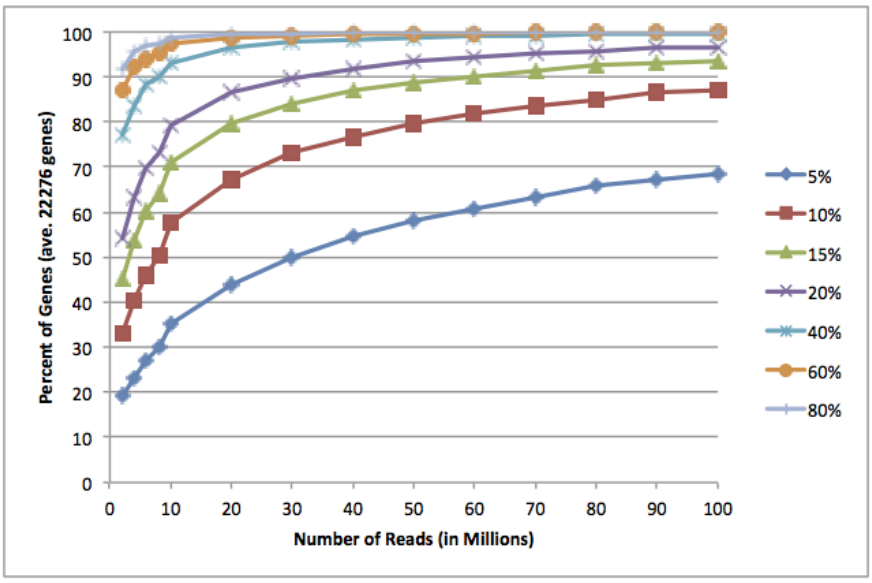


Fig. 4.S6. Effect of sequencing depth on expression values Bins containing mapped reads were created by randomly sampling mapped reads equally from each of the eight tissues, to reach final read counts of 2, 4, 6, 8, 10, 20, 30, 40, 60, 70, 80, 90, and 100 million reads. Read counts for each of our AUGUSTUS gene models were generated using the htseq-

count command from HTseq as previously described, and we generated expression values in FPKM for each of our bins.

Table 4.S1 (online version of article)

Gene expression in electric fishes. Combined table of gene expression the eight tissues (brain, spinal cord, heart, skeletal muscle, kidney, main EO, Sachs' EO, Hunter's EO) of *E. electricus,* (b) Gene expression in skeletal muscle and EO of *S. macrurus*, (c) Gene expression in skeletal muscle and EO of *E. virescens*, (d) Gene expression in skeletal muscle and EO samples of *B. brachyistius*.

Table 4.S2 (online version of article)

Relative expression abundance of muscle genes in electric organ relative to skeletal muscle in four species of electric fish. Colors indicate overexpression (red) and underexpression (green) in electric organ relative to skeletal muscle. n-d indicates transcripts not found or ratios not determined because abundance was too low for accurate quantitation.

Table 4.S3. Summary of literature regarding transcription factor interactions in figure 4.2b. For each gene, gene name, whether it is expressed in presomatic mesoderm/somites is listed, along with citation, whether it is expressed in mature muscle along with citation, whether expression is known to inhibit muscle development along with citation, and various notes about expression patterns and interactions are listed.

Table 4.S4. Summary of literature regarding protein interactions and locations in figure 4.2c. For each protein (where known), its location in skeletal muscle, known binding partners, results of knockout experiments in other vertebrates, and roles in myopathy are listed. Notes summarize other potentially relevant features of these proteins.

Gene Name	Location	Binding Partner	Phenotype of KO/Mutation	Myopathy	Notes
smyd1	M line (*14*, *76*)	myosin (*14*, *76*)	disrupts sarcomere (*14*, *76*, *77*)		Smyd proteins stabilize sarcomeric proteins (*76-79*). Smyd1 is activated by myogenin (*79*, *80*)
smyd2b	I band (*78*, *79*)	titin (*78*)	disrupts sarcomere (*78*)		Smyd proteins stabilize sarcomeric proteins (*76-79*)
hspb11			disrupts Sarcomere (*14*)		
fbxo40		IRSI (*81*)	enlarged myofibers (*81*)		Fbxo40 complex limits IGF1 signaling by inducing IRIS1 ubiquination (*81*). The activated IRIS1/IGFR complex inhibits protein degradation by inhibiting Foxo (*21*)
IGF2b	secreted by cell in autocrine fashion (*19*, *80*)	IGF1 receptor (*80*)	muscle atrophy (*80*) severe hypoplasia (*82*)	Attenuated IGF signaling associated with cardiac myopathy (*83*).	Enhances body size and developmental rate in zebrafish (*84*). Overexpression prevents muscle atrophy in mice (*85*). IGF activates IRIS1, IRIS1/IGFR complex binds to PIK3 (*20*). This activated complex can influence cell proliferation and increase protein synthesis by inhibiting Foxo (*21*)

(4.S4 continued)

ARHGEF12	cytoplasm (*86*)	IGF1 receptor (*86*)			
pik3r3b	cytoplasm (*87*)	IRS1 (*87*)			
NET37-like	nuclear envelope (*23*)	IGF2 (*23*)	prevents muscle differentiation, decreases myogenin (*23*)		Increases autocrine release of IGFII (*23*)
col6a6	connective tissue, basal lamina (*88*)	col6a3 (*88*)		Muscular Dystrophy (*88*)	
col14a1	connective tissue, basal lamina (*89*)	type I collagen (*89*)	prevents glycosylation of dystroglycan (*89*)		
dmd	plasma membrane (*90*)	dystroglycan (*90*)		muscular dystrophy (*90*)	
glytl1b	Golgi apparatus (*91*)	alpha dystroglycan (*91*)		muscular dystrophy (*91*)	

Table 4.S5. Summary of genomic sequencing experiments. Genomic DNA from *E. electricus* was sequenced using Illumina technologies.

Platform	library type	median read length (nt)	number of reads (millions)
Illumina GAIIx	paired-end	73	37
Illumina HiSeq 2000	paired-end	99	600
Illumina HiSeq 2000	2k mate-pair	51	830

Table 4.S6. Summary of transcriptomic sequencing experiments. RNA-Seq experiments were performed in *E. electricus, S. macrurus, E. virescens, M. electricus* and *B. brachyistius*.

Species	samples	sequencing technology	run type
E. electricus	3 tissues - main EO, Sachs' EO, skeletal muscle	Illumina GAIIx	paired end
E. electricus	7 tissue - brain, spinal cord, main EO, Sachs' EO, Hunter's EO, skeletal muscle, heart	Illumina HiSeq 2000	single end
E. electricus	1 tissue - kidney	Illumina HiSeq 2000	paired end
S. macrurus	2 tissues - EO and skeletal muscle	Illumina HiSeq 2000	paired end
E. virescens	2 tissues - EO and skeletal muscle	Illumina HiSeq 2000	paired end
E. virescens	2 tissues - EO and skeletal muscle	Illumina HiSeq 2000	paired end
B. brachyistius	2 tissues - EO and skeletal muscle	Illumina HiSeq 2000	paired end
M. electricus	2 tissues - EO and skeletal muscle	Illumina HiSeq 2000	paired end

Table 4.S7. Summary of *E. electricus g*enome assembly. Summary statistics were calculated for (a) both gapped (containing Ns) and ungapped (Ns stripped) builds of SOAPdenovo2 assembly.

	Gapped	Ungapped
Number of Contigs	121323	121323
Total Length	560202223	523576209
Average length	4617.4	4315.6
Longest Contig	996512	972603
Shortest Contig	100	100
Contig N50	104253	103026
Contig N90	12699	15371
G/C Content	42.50%	42.50%

References

1. J. S. Albert, W. G. R. Crampton, “Electroreception and electrogenesis.” in *The Physiology of Fishes* (Springer, New York, ed. 3, 2005), pp. 431–472.
2. C. Darwin, in *On the Origin of Species by Means of Natural Selection* (J. Murray, London, 1859), pp. ix, 1.
3. B. A. Block, Evolutionary novelties: How fish have built a heater out of muscle. *Am. Zool.* **31**, 726–742 (1991).
4. H. Kokubo, S. Tomita-Miyagawa, Y. Hamada, Y. Saga, Hesr1 and Hesr2 regulate atrioventricular boundary formation in the developing heart through the repression of Tbx2. *Development* **134**, 747–755 (2007). Medline doi:10.1242/dev.02777
5. G. A. Unguez, H. H. Zakon, Reexpression of myogenic proteins in mature electric organ after removal of neural input. *J. Neurosci.* **18**, 9924–9935 (1998). Medline
6. Materials and methods are available as supplementary materials on *Science* Online.
7. H. Ghanbari, H. C. Seo, A. Fjose, A. W. Brändli, Molecular cloning and embryonic expression of *Xenopus Six* homeobox genes. *Mech. Dev.* **101**, 271–277 (2001). Medline doi:10.1016/S0925-4773(00)00572-4
8. F. Spitz, J. Demignon, A. Porteu, A. Kahn, J. P. Concordet, D. Daegelen, P. Maire, Expression of myogenin during embryogenesis is controlled by Six/*sine oculis* homeoproteins through a conserved MEF3 binding site. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 14220–14225 (1998). Medline doi:10.1073/pnas.95.24.14220
9. K. Kawakami, S. Noguchi, M. Noda, H. Takahashi, T. Ohta, M. Kawamura, H. Nojima, K. Nagano, T. Hirose, S. Inayama, H. Hayashida, T. Miyata, S. Numa, Primary structure of

the α-subunit of *Torpedo californica* (Na+ + K+)ATPase deduced from cDNA sequence. *Nature* **316**, 733–736 (1985). Medline doi:10.1038/316733a0

10. M. F. Buas, S. Kabak, T. Kadesch, The Notch effector Hey1 associates with myogenic target genes to repress myogenesis. *J. Biol. Chem.* **285**, 1249–1258 (2010). Medline doi:10.1074/jbc.M109.046441

11. K. J. Nowak, K. E. Davies, Duchenne muscular dystrophy and dystrophin: Pathogenesis and opportunities for treatment. *EMBO Rep.* **5**, 872–876 (2004). Medline doi:10.1038/sj.embor.7400221

12. R. G. Berry, S. Despa, W. Fuller, D. M. Bers, M. J. Shattock, Differential distribution and regulation of mouse cardiac Na+/K+-ATPase α1 and α2 subunits in T-tubule and surface sarcolemmal membranes. *Cardiovasc. Res.* **73**, 92–100 (2007). Medline doi:10.1016/j.cardiores.2006.11.006

13. J. Lowe, G. M. Araujo, A. R. Pedrenho, N. Nunes-Tavares, M. G. Ribeiro, A. Hassón-Voloch, Polarized distribution of Na+, K+-ATPase α-subunit isoforms in electrocyte membranes. *Biochim. Biophys. Acta* **1661**, 40–46 (2004). Medline doi:10.1016/j.bbamem.2003.11.020

14. N. Klüver, L. Yang, W. Busch, K. Scheffler, P. Renner, U. Strähle, S. Scholz, Transcriptional response of zebrafish embryos exposed to neurotoxic compounds reveals a muscle activity dependent *hspb11* expression. *PLOS ONE* **6**, e29063 (2011). Medline doi:10.1371/journal.pone.0029063

15. C. Duan, H. Ren, S. Gao, Insulin-like growth factors (IGFs), IGF receptors, and IGF-binding proteins: Roles in skeletal muscle growth and differentiation. *Gen. Comp. Endocrinol.* **167**, 344–351 (2010). Medline doi:10.1016/j.ygcen.2010.04.009

16. B. C. Hoopes, M. Rimbault, D. Liebers, E. A. Ostrander, N. B. Sutter, The insulin-like growth factor 1 receptor (IGF1R) contributes to reduced size in dogs. *Mamm. Genome* **23**, 780–790 (2012). Medline doi:10.1007/s00335-012-9417-z

17. N. B. Sutter, C. D. Bustamante, K. Chase, M. M. Gray, K. Zhao, L. Zhu, B. Padhukasahasram, E. Karlins, S. Davis, P. G. Jones, P. Quignon, G. S. Johnson, H. G. Parker, N. Fretwell, D. S. Mosher, D. F. Lawler, E. Satyaraj, M. Nordborg, K. G. Lark, R. K. Wayne, E. A. Ostrander, A single *IGF1* allele is a major determinant of small size in dogs. *Science* **316**, 112–115 (2007). Medline doi:10.1126/science.1137045

18. D. J. Emlen, I. A. Warren, A. Johns, I. Dworkin, L. C. Lavine, A mechanism of extreme growth and reliable signaling in sexually selected ornaments and weapons. *Science* **337**, 860–864 (2012). Medline doi:10.1126/science.1224286

19. A. S. Van Laere, M. Nguyen, M. Braunschweig, C. Nezer, C. Collette, L. Moreau, A. L. Archibald, C. S. Haley, N. Buys, M. Tally, G. Andersson, M. Georges, L. Andersson, A regulatory mutation in *IGF2* causes a major QTL effect on muscle growth in the pig. *Nature* **425**, 832–836 (2003). Medline doi:10.1038/nature02064

20. I. Mothe, L. Delahaye, C. Filloux, S. Pons, M. F. White, E. Van Obberghen, Interaction of wild type and dominant-negative p55PIK regulatory subunit of phosphatidylinositol 3-kinase with insulin-like growth factor-1 signaling proteins. *Mol. Endocrinol.* **11**, 1911–1923 (1997). Medline doi:10.1210/mend.11.13.0029

21. A. Otto, K. Patel, Signalling and the control of skeletal muscle size. *Exp. Cell Res.* **316**, 3059–3066 (2010). Medline doi:10.1016/j.yexcr.2010.04.009

22. I. H. Chen, M. Huber, T. Guan, A. Bubeck, L. Gerace, Nuclear envelope transmembrane proteins (NETs) that are up-regulated during myogenesis. *BMC Cell Biol.* **7**, 38 (2006). Medline doi:10.1186/1471-2121-7-38

23. K. Datta, T. Guan, L. Gerace, NET37, a nuclear envelope transmembrane protein with glycosidase homology, is involved in myoblast differentiation. *J. Biol. Chem.* **284**, 29666–29676 (2009). Medline doi:10.1074/jbc.M109.034041

24. S. E. Mate, K. J. Brown, E. P. Hoffman, Integrated genomics and proteomics of the *Torpedo californica* electric organ: Concordance with the mammalian neuromuscular junction. *Skeletal Muscle* **1**, 20 (2011). Medline doi:10.1186/2044-5040-1-20

25. M. E. Alfaro, F. Santini, C. Brock, H. Alamillo, A. Dornburg, D. L. Rabosky, G. Carnevale, L. J. Harmon, Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 13410–13414 (2009). Medline doi:10.1073/pnas.0811087106

26. C. Linneaus, *Systema Naturae* (Holmiae, ed. XII, 1766).

27. J. S. Albert, *Check List of the Freshwater Fishes of South and Central America* (Edipuers, Porto Alegre, Brazil, 2003).

28. A. Valenciennes, “Les poissons.” in *Le Règne Animal Distribué d'Après son Organisation, pour Servir de Base à l'Historie Naturelle des Animaux, et d'Introduction a l'Anatomie Comparée*, G. Cuvier, Ed. (Masson, Paris, ed. 3, 1838–1842).

29. K. R. Bradnam, J. N. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert, J. A. Chapman, G. Chapuis, R. Chikhi, H. Chitsaz, W.-C. Chou, J. Corbeil, C. Del Fabbro, T. R. Docking, R. Durbin, D. Earl, S. Emrich, P. Fedotov, N. A. Fonseca, G. Ganapathy, R.

A. Gibbs, S. Gnerre, É. Godzaridis, S. Goldstein, M. Haimel, G. Hall, D. Haussler, J. B. Hiatt, I. Y. Ho, J. Howard, M. Hunt, S. D. Jackman, D. B. Jaffe, E. Jarvis, H. Jiang, S. Kazakov, P. J. Kersey, J. O. Kitzman, J. R. Knight, S. Koren, T.-W. Lam, D. Lavenier, F. Laviolette, Y. Li, Z. Li, B. Liu, Y. Liu, R. Luo, I. MacCallum, M. D. MacManes, N. Maillet, S. Melnikov, B. M. Vieira, D. Naquin, Z. Ning, T. D. Otto, B. Paten, O. S. Paulo, A. M. Phillippy, F. Pina-Martins, M. Place, D. Przybylski, X. Qin, C. Qu, F. J. Ribeiro, S. Richards, D. S. Rokhsar, J. G. Ruby, S. Scalabrin, M. C. Schatz, D. C. Schwartz, A. Sergushichev, T. Sharpe, T. I. Shaw, J. Shendure, Y. Shi, J. T. Simpson, H. Song, F. Tsarev, F. Vezzi, R. Vicedomini, J. Wang, K. C. Worley, S. Yin, S.-M. Yiu, J. Yuan, G. Zhang, H. Zhang, S. Zhou, I. F. Korf, Assemblathon 2: Evaluating de novo methods of genome assembly in three vertebrate species. http://arxiv.org/abs/1301.5406 (2013).

30. J. F. Gmelin, *Systema Naturae per Regna Tria Naturae, Secundum Classes, Ordines, Genera, Species; cum Characteribus, Differentiis, Synonymis, Locis*. Editio decimo tertia, aucta, reformata. 3 vols. in 9 parts. (Lipsiae, 1788–1993), vol. 1 (pt 3), pp. 1033–1516.

31. T. Gregory, Animal Genome Size Database (2013); www.genomesize.com.

32. D. C. Hardie, P. D. N. Hebert, The nucleotypic effects of cellular DNA content in cartilaginous and ray-finned fishes. *Genome* **46**, 683–706 (2003). Medline doi:10.1139/g03-040

33. P. Chomczynski, N. Sacchi, The single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction: Twenty-something years on. *Nat. Protoc.* **1**, 581–585 (2006). Medline doi:10.1038/nprot.2006.83

34. D. C. Hardie, P. D. N. Hebert, Genome-size evolution in fishes. *Can. J. Fish. Aquat. Sci.* **61**, 1636–1646 (2004). doi:10.1139/f04-106

35. R. Hinegardner, Evolution of cellular DNA content in teleost fishes. *Am. Nat.* **102**, 517–523 (1968). doi:10.1086/282564

36. R. Hinegardner, D. E. Rosen, Cellular DNA content and the evolution of teleostean fishes. *Am. Nat.* **106**, 621–644 (1972). doi:10.1086/282801

37. J. T. Simpson, Exploring genome characteristics and sequence quality without a reference. http://arxiv.org/abs/1307.8026 (2013).

38. A. E. Vinogradov, Genome size and GC-percent in vertebrates as determined by flow cytometry: The triangular relationship. *Cytometry* **31**, 100–109 (1998). Medline doi:10.1002/(SICI)1097-0320(19980201)31:2<100::AID-CYTO5>3.0.CO;2-Q

39. R. Luo, B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu, C. Han, D. W. Cheung, S. M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T. W. Lam, J. Wang, SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18 (2012). Medline doi:10.1186/2047-217X-1-18

40. T. D. Wu, C. K. Watanabe, GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005). Medline doi:10.1093/bioinformatics/bti310

41. G. Parra, K. Bradnam, I. Korf, CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007). Medline doi:10.1093/bioinformatics/btm071

42. C. Trapnell, L. Pachter, S. L. Salzberg, TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009). Medline doi:10.1093/bioinformatics/btp120

43. M. Stanke, B. Morgenstern, AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33** (suppl. 2), W465–W467 (2005). Medline doi:10.1093/nar/gki458

44. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990). Medline doi:10.1016/S0022-2836(05)80360-2

45. P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A. K. Kähäri, D. Keefe, S. Keenan, R. Kinsella, M. Komorowska, G. Koscielny, E. Kulesha, P. Larsson, I. Longden, W. McLaren, M. Muffato, B. Overduin, M. Pignatelli, B. Pritchard, H. S. Riat, G. R. Ritchie, M. Ruffier, M. Schuster, D. Sobral, Y. A. Tang, K. Taylor, S. Trevanion, J. Vandrovcova, S. White, M. Wilson, S. P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. M. Fernández-Suarez, J. Harrow, J. Herrero, T. J. Hubbard, A. Parker, G. Proctor, G. Spudich, J. Vogel, A. Yates, A. Zadissa, S. M. Searle, Ensembl 2012. *Nucleic Acids Res.* **40**, D84–D90 (2012). Medline doi:10.1093/nar/gkr991

46. M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev,

Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011). Medline doi:10.1038/nbt.1883

47. S. Anders, P. T. Pyl, W. Huber, HTSeq – A Python framework to work with high-throughput sequencing data. http://biorxiv.org/content/early/2014/02/20/002824 (2014); doi: 10.1101/002824.

48. S. Anders, W. Huber, Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010). Medline doi:10.1186/gb-2010-11-10-r106

49. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009). Medline doi:10.1186/gb-2009-10-3-r25

50. B. Li, C. N. Dewey, RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011). Medline doi:10.1186/1471-2105-12-323

51. D. Tritchler, E. Parkhomenko, J. Beyene, Filtering genes for cluster and network analysis. *BMC Bioinformatics* **10**, 193 (2009). Medline doi:10.1186/1471-2105-10-193

52. R Development Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2008); www.R-project.org.

53. M. Gouy, S. Guindon, O. Gascuel, SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* **27**, 221–224 (2010). Medline doi:10.1093/molbev/msp259

54. L. Al-Qusairi, J. Laporte, T-tubule biogenesis and triad formation in skeletal muscle and implication in human diseases. *Skeletal Muscle* **1**, 26 (2011). Medline doi:10.1186/2044-5040-1-26

55. S. Lange, E. Ehler, M. Gautel, From A to Z and back? Multicompartment proteins in the sarcomere. *Trends Cell Biol.* **16**, 11–18 (2006). Medline doi:10.1016/j.tcb.2005.11.007

56. S. Treves, M. Vukcevic, M. Maj, R. Thurnheer, B. Mosca, F. Zorzato, Minor sarcoplasmic reticulum membrane components that modulate excitation-contraction coupling in striated muscles. *J. Physiol.* **587**, 3071–3079 (2009). Medline doi:10.1113/jphysiol.2009.171876

57. M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, D. G. Higgins, Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007). Medline doi:10.1093/bioinformatics/btm404

58. A. H. Bass, in *Electroreception* (Wiley, New York, 1986), pp. 13–70.

59. J. R. Gallant, C. D. Hopkins, D. L. Deitcher, Differential expression of genes and proteins between electric organ and skeletal muscle in the mormyrid electric fish *Brienomyrus brachyistius*. *J. Exp. Biol.* **215**, 2479–2494 (2012). Medline doi:10.1242/jeb.063222

60. I. R. Schwartz, G. D. Pappas, M. V. Bennett, The fine structure of electrocytes in weakly electric teleosts. *J. Neurocytol.* **4**, 87–114 (1975). Medline doi:10.1007/BF01099098

61. G. A. Unguez, H. H. Zakon, Phenotypic conversion of distinct muscle fiber populations to electrocytes in a weakly electric fish. *J. Comp. Neurol.* **399**, 20–34 (1998). Medline doi:10.1002/(SICI)1096-9861(19980914)399:1<20::AID-CNE2>3.0.CO;2-C

62. M. A. Esquibel, I. Alonso, H. Meyer, C. Chagas, G. de Castro, [Some aspects of the histogenesis and ontogenesis of electric organs in *Electrophorus electricus* (L.)]. *C. R. Acad. Sci. Hebd. Seances Acad. Sci. D* **273**, 196–199 (1971). Medline

63. R. D. Machado, W. de Souza, G. Cotta-Pereira, G. de Oliveira Castro, On the fine structure of the electrocyte of *Electrophorus* electricus L. *Cell Tissue Res.* **174**, 355–366 (1976). Medline doi:10.1007/BF00220681

64. C. Winkler, H. Elmasri, B. Klamt, J. N. Volff, M. Gessler, Characterization of hey bHLH genes in teleost fish. *Dev. Genes Evol.* **213**, 541–553 (2003). Medline doi:10.1007/s00427-003-0360-6

65. L. Sun, H. Xie, M. A. Mori, R. Alexander, B. Yuan, S. M. Hattangadi, Q. Liu, C. R. Kahn, H. F. Lodish, Mir193b-365 is essential for brown fat differentiation. *Nat. Cell Biol.* **13**, 958–965 (2011). Medline doi:10.1038/ncb2286

66. J. Sun, C. N. Kamei, M. D. Layne, M. K. Jain, J. K. Liao, M. E. Lee, M. T. Chin, Regulation of myogenic terminal differentiation by the hairy-related transcription factor CHF2. *J. Biol. Chem.* **276**, 18591–18596 (2001). Medline doi:10.1074/jbc.M101163200

67. G. E. Muscat, J. Emery, E. S. Collie, Tissue-specific expression of the skeletal alpha-actin gene involves sequences that can function independently of MyoD and Id. *Gene Expr*. **2**, 241–257 (1992). Medline

68. J. D. Meissner, P. K. Umeda, K. C. Chang, G. Gros, R. J. Scheibe, Activation of the beta myosin heavy chain promoter by MEF-2D, MyoD, p300, and the calcineurin/NFATc1 pathway. *J. Cell. Physiol*. **211**, 138–148 (2007). Medline doi:10.1002/jcp.20916

69. D. Li, Z. Niu, W. Yu, Y. Qian, Q. Wang, Q. Li, Z. Yi, J. Luo, X. Wu, Y. Wang, R. J. Schwartz, M. Liu, SMYD1, the myogenic activator, is a direct target of serum response factor and myogenin. *Nucleic Acids Res*. **37**, 7059–7071 (2009). Medline doi:10.1093/nar/gkp773

70. G. Oliver, A. Mailhos, R. Wehr, N. G. Copeland, N. A. Jenkins, P. Gruss, Six3, a murine homologue of the sine oculis gene, demarcates the most anterior border of the developing neural plate and is expressed during eye development. *Development* **121**, 4045–4055 (1995). Medline

71. K. Kawakami, H. Ohto, K. Ikeda, R. G. Roeder, Structure, function and expression of a murine homeobox protein AREC3, a homologue of Drosophila sine oculis gene product, and implication in development. *Nucleic Acids Res.* **24**, 303–310 (1996). Medline doi:10.1093/nar/24.2.303

72. B. Thisse, C. Thisse, Fast release clones: A high throughput expression analysis. ZFIN Direct Data Submission (2004); http://zfin.org.

73. D. A. Bessarab, S. W. Chong, B. P. Srinivas, V. Korzh, Six1a is required for the onset of fast muscle differentiation in zebrafish. *Dev. Biol.* **323**, 216–228 (2008). Medline doi:10.1016/j.ydbio.2008.08.015

74. A. F. Richard, J. Demignon, I. Sakakibara, J. Pujol, M. Favier, L. Strochlic, F. Le Grand, N. Sgarioto, A. Guernec, A. Schmitt, N. Cagnard, R. Huang, C. Legay, I. Guillet-Deniau, P. Maire, Genesis of muscle fiber-type diversity during mouse embryogenesis relies

on Six1 and Six4 gene expression. *Dev. Biol.* **359**, 303–320 (2011). Medline doi:10.1016/j.ydbio.2011.08.010

75. M. Kobayashi, H. Osanai, K. Kawakami, M. Yamamoto, Expression of three zebrafish Six4 genes in the cranial sensory placodes and the developing somites. Mech. Dev. **98**, 151–155 (2000). Medline doi:10.1016/S0925-4773(00)00451-2

76. S. Just, B. Meder, I. M. Berger, C. Etard, N. Trano, E. Patzel, D. Hassel, S. Marquart, T. Dahme, B. Vogel, M. C. Fishman, H. A. Katus, U. Strähle, W. Rottbauer, The myosin-interacting protein SMYD1 is essential for sarcomere organization. *J. Cell Sci.* **124**, 3127–3136 (2011). Medline doi:10.1242/jcs.084772

77. X. Tan, J. Rotllant, H. Li, P. De Deyne, S. J. Du, SmyD1, a histone methyltransferase, is required for myofibril organization and muscle contraction in zebrafish embryos. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 2713–2718 (2006). Medline doi:10.1073/pnas.0509503103

78. L. T. Donlin, C. Andresen, S. Just, E. Rudensky, C. T. Pappas, M. Kruger, E. Y. Jacobs, A. Unger, A. Zieseniss, M. W. Dobenecker, T. Voelkel, B. T. Chait, C. C. Gregorio, W. Rottbauer, A. Tarakhovsky, W. A. Linke, Smyd2 controls cytoplasmic lysine methylation of Hsp90 and myofilament organization. *Genes Dev.* **26**, 114–119 (2012). Medline doi:10.1101/gad.177758.111

79. H. Li, Y. Zhong, Z. Wang, J. Gao, J. Xu, W. Chu, J. Zhang, S. Fang, S. J. Du, Smyd1b is required for skeletal and cardiac muscle function in zebrafish. Mol. Biol. Cell **24**, 3511–3521 (2013). Medline doi:10.1091/mbc.E13-06-0352

80. S. Schiaffino, C. Mammucari, Regulation of skeletal muscle growth by the IGF1-Akt/PKB pathway: Insights from genetic models. *Skeletal Muscle* **1**, 4 (2011). Medline doi:10.1186/2044-5040-1-4

81. J. Shi, L. Luo, J. Eash, C. Ibebunjo, D. J. Glass, The SCF-Fbxo40 complex induces IRS1 ubiquitination in skeletal muscle, limiting IGF1 signaling. *Dev. Cell* **21**, 835–847 (2011). Medline doi:10.1016/j.devcel.2011.09.011

82. J. P. Liu, J. Baker, A. S. Perkins, E. J. Robertson, A. Efstratiadis, Mice carrying null mutations of the genes encoding insulin-like growth factor I (Igf-1) and type 1 IGF receptor (Igf1r). *Cell* **75**, 59–72 (1993). Medline

83. L. Elia, R. Contu, M. Quintavalle, F. Varrone, C. Chimenti, M. A. Russo, V. Cimino, L. De Marinis, A. Frustaci, D. Catalucci, G. Condorelli, Reciprocal regulation of microRNA-1 and insulin-like growth factor-1 signal transduction cascade in cardiac and skeletal muscle in physiological and pathological conditions. *Circulation* **120**, 2377–2385 (2009). Medline doi:10.1161/CIRCULATIONAHA.109.879429

84. X. Wang, L. Lu, Y. Li, M. Li, C. Chen, Q. Feng, C. Zhang, C. Duan, Molecular and functional characterization of two distinct IGF binding protein-6 genes in zebrafish. Am. J. Physiol. Regul. Integr. *Comp. Physiol*. **296**, R1348–R1357 (2009). Medline doi:10.1152/ajpregu.90969.2008

85. E. R. Barton-Davis, D. I. Shoturma, A. Musaro, N. Rosenthal, H. L. Sweeney, Viral mediated expression of insulin-like growth factor I blocks the aging-related loss of skeletal muscle function. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 15603–15607 (1998). Medline doi:10.1073/pnas.95.26.15603

86. S. Taya, N. Inagaki, H. Sengiku, H. Makino, A. Iwamatsu, I. Urakawa, K. Nagao, S. Kataoka, K. Kaibuchi, Direct interaction of insulin-like growth factor-1 receptor with leukemia-associated RhoGEF. *J. Cell Biol.* **155**, 809–820 (2001). Medline doi:10.1083/jcb.200106139

87. P. R. Shepherd, Mechanisms regulating phosphoinositide 3-kinase signalling in insulin-sensitive tissues. *Acta Physiol. Scand*. **183**, 3–12 (2005). Medline doi:10.1111/j.1365-201X.2004.01382.x

88. W. R. Telfer, A. S. Busta, C. G. Bonnemann, E. L. Feldman, J. J. Dowling, Zebrafish models of collagen VI-related myopathies. *Hum. Mol. Genet*. **19**, 2433–2444 (2010). Medline doi:10.1093/hmg/ddq126

89. H. L. Bader, E. Lambert, A. Guiraud, M. Malbouyres, W. Driever, M. Koch, F. Ruggiero, Zebrafish collagen XIV is transiently expressed in epithelia and is required for proper function of certain basement membranes. *J. Biol. Chem*. **288**, 6777–6787 (2013). Medline doi:10.1074/jbc.M112.430637

90. J. Berger, P. D. Currie, Zebrafish models flex their muscles to shed light on muscular dystrophies. *Dis. Model*. Mech. **5**, 726–732 (2012). Medline doi:10.1242/dmm.010082

91. M. Brockington, S. Torelli, P. Prandini, C. Boito, N. F. Dolatshad, C. Longman, S. C. Brown, F. Muntoni, Localization and functional analysis of the LARGE family of glycosyltransferases: Significance for muscular dystrophy. Hum. Mol. Genet. **14**, 657–665 (2005). Medline doi:10.1093/hmg/ddi062

Chapter 5: A tail of two voltages: quantitative proteomic comparison of the three electric organs of *E. electricus.*

Lindsay L. Traeger[1,2], Grzegorz Sabat[2], Gregory A. Barrett-Wilt[2], Michael R. Sussman[2,3]

Affiliations:

[1] Genetics Department, University of Wisconsin-Madison, Madison, WI 53706, USA.

[2] Biotechnology Center, University of Wisconsin-Madison, Madison, WI, 53706, USA.

[3] Department of Biochemistry, University of Wisconsin-Madison, Madison, WI, 53706, USA.

This chapter represents a manuscript that is in progress.

Abstract

Ion channels and pumps are essential for the function of all nerve and muscle cells, but in addition, play a unique role during the electrogenesis observed exclusively in electric fish. The strong voltage electric eel is unique amongst electric fishes because it has three distinct electric organs that simultaneously allow it to emit continuous pulses of electric discharge for communication and navigation as well as the very strong intermittent discharge used for defense and predation. Recent genome sequencing efforts with the strong-voltage electric eel have now provided the opportunity to perform mass spectrometric-based quantitative proteomic studies, and I describe here the use of isotope-assisted multiplexed technology to compare the proteome and phosphoproteome of the three electric organs. These experiments provide a means to test the hypothesis that differences in the proteome and phosphoproteome of these three organs underlie their distinct energetic needs. I have discovered novel phosphorylation sites in the sodium pump and sodium channels and identified a potassium channel that shows a particularly unique difference in protein concentration among the three organs. Additionally, I find a number of transcription factors and protein kinases that are differentially abundant in main and Sachs' EO. Together, these findings are consistent with the hypothesis that the differences in proteomes of these three organs underlie unique differences in the energetic needs of each cell type, reflecting the trade off between generating weak voltages continuously and strong voltages intermittently.

Introduction.

The electric eel (*Electrophorus electricus*) is a fresh water electric fish native to South America, most famous for its remarkable ability to produce high-voltage electric discharges for predation and defense. The evolutionary history of electrogenic organs in fishes is particularly interesting, as not only have electric organs (EOs) evolved in multiple independent fish lineages from myogenic precursors [1, 2], but within each lineage, there is a tremendous amount of variation. This is particularly true in Gymnotiformes where, for example, *E. electricus* is the only species in Gymnotiformes identified to date that has three distinct EOs and is also capable of generating strong-voltage discharges [3]. Like all other Gymnotiformes, *E. electricus* is able to generate weak electric organ discharges (EODs) (Sachs' EO, and the anterior portion of the Hunter's EO) for the purposes of navigation and communication in the murky South American fresh waters in which they live. However, in addition to this electro navigation and electro communicative function, *E. electricus* can produce enormous EODs in excess of 600 volts and one ampere of current [4](main EO and posterior portion of Hunter's EO) for the purposes of predation, defense, and "remote controlling" prey—forcing hiding prey to reveal their location by using high-voltage discharges to induce muscle contractions [5]. Because of this remarkable ability to produce discharges with very high electric energy, *E. electricus* and other strong-voltage species such as *Torpedo californica* are incredibly rich sources for proteins involved in EOD, and as such, were an important source of these proteins in early biochemical investigations on the structure and function of ion channels and ion pumps [6].

Differences in tissue physiology in organisms arise from differences in both protein abundance, and differences in protein regulation, such as by posttranslational modifications, including the phosphorylation of serine, tyrosine and threonine residues in

proteins. Mass spectrometry approaches can quantitatively measure both of these aspects. Further, post protein extraction and digestion, peptides from different samples can be given a unique isotope-assisted isobaric tag (e.g., ten-plex with Tandem Mass Tag, or TMT) covalently attached to peptides in different independent samples, allowing users to combine samples post-labeling in equal ratios, perform further processing on all biological samples together, and analyze simultaneously in the mass spectrometer [7].

Recently, we used a comparative genomic and transcriptomic approach to understand the evolution of EOs across lineages of fishes that have independently evolved EOs, and showed that several transcription factors and signaling pathways showed similar expression patterns in EOs despite independent evolutionary origins [8]. The strong voltage electric eel (*E. electricus*) is unique since it has three distinct EOs, and little or nothing is known about their comparative structure at the molecular level. Here, I present the first quantitative proteomic and phosphoproteomic comparison of the three *E. electricus* EOs and skeletal muscle. I reveal differences in protein abundance and phosphorylation that I hypothesize may underlie the functional differences of individual organs, e.g. in their production of weak voltage continuously for navigation and communication or strong voltages intermittently for predation or defense. Second, I observed differences in the three EOs, in protein abundance and phosphorylation for proteins of unknown function in EOs. Together, in contrast to earlier proteomics studies in *Torpedo* examining a few hundred EO proteins without a reference genome of the same species [9, 10], this untargeted, isotope-assisted quantification of proteins and phosphoproteins in the electric organs represent the first bona fide, large-scale proteomic

comparison of the EOs and muscle in any electric fish, and provide the baseline infrastructure for all future genomic efforts on understanding electrocyte biology.

Results/Discussion

The essential sodium transport proteins involved in electrogenesis have novel phosphorylation sites.

Being organs uniquely dedicated to generating exogenous voltages, it is not surprising to observe that the most abundant proteins in these organs are those vital to initiating EOD (acetylcholine receptors, a Na^+ channel), propagating the action potential (voltage-gated Na^+ channels), and returning cells to and maintaining homeostasis (acetylcholine esterase, Na^+/K^+ ATPase, K^+ channels). Further, the polarized arrangement of electrocytes within EOs is critical for the individual voltages generated in each electrocyte to summate (reviewed in [11], Figure 5.1A).

Each electrocyte is a large, flattened cell with two major membrane faces: the innervated membrane face, which receives nervous stimulation, contains numerous sodium channels and opposite to it, the non-innervated membrane face, whish is highly invaginated, contains numerous Na^+/K^+-ATPase. Upon activation by acetylcholine release from the innervating neuron, the acetylcholine receptors open, allowing Na^+ to flow into the cell and causing a slight membrane depolarization; in turn, this slight depolarization of the membrane causes the voltage-gated Na^+ channels to open, and a massive Na^+ influx in a polarized manner across the innervated membrane face occurs, driving the innervated membrane face into positive potentials. Because the non-innervated membrane face remains at negative potentials due to the abundance of Na^+/K^+-ATPase present in the

membrane, there is a large transcellular potential difference that is achieved within each electrocyte, and because electrocytes are arranged massively in series and in parallel, voltages generated in each electrocyte summate. It is the function of Na^+/K^+-ATPase and potassium channels to return the membranes to the 'resting' membrane potential

Given the sheer abundance of these electrogenic proteins in EO, it is not surprising that I found numerous phosphorylated residues in them (Figure 5.1A and Table 5.S11). In total, for ATP1a2a I identified 16 phosphorylated residues in ATP1a2a, four of which have not been described previously in mammals [12], SCN4aa (13 novel and one known phosphosite), SCN4ab (five novel phosphosites), the acetylcholine receptor subunits CHRNB1 (one novel and one known phosphosite), CHRND (two novel and two known phosphosites), CHRNE (two novel phosphosites), and acetylcholine esterase (two novel phosphosites). I was unable to detect any phosphosites for K^+ channels.

Further studies of the impact of these phosphorylation sites, perhaps in a heterologous system or *ex vivo* using agonists for various channels or pumps in electric organ slices (see Appendix I), will aid in understanding the function of these novel sites

<u>SCN4aa phosphorylation, FGF13, and calmodulin: C-terminal regulatory mechanism for the voltage gated Na+ channels of Gymnotiformes?</u>

Calmodulin is an extremely abundant protein in *E. electricus* electric organ, making up to two percent of the total protein content of the cell by some estimates [13], and is among the top 10% most abundant proteins in main EO and Hunter's EO, but not Sachs' EO or muscle (see Table 5.S6). Despite its abundance, a clear role for calmodulin in EO function has not been described. It was recently demonstrated that the voltage gated Na^+ channel,

SCN4a, is regulated by Ca^{2+} and calmodulin through interactions of calmodulin with the C-terminal domain of SCN4a [14]. Further, the crystal structure for the ternary complex of a voltage gated Na^{+} channel, calmodulin, and fibroblast growth factor 13 (FGF13) was described, highlighting a role for FGF13 in voltage-gated Na^{+} channel regulation via the C-terminal domain [15]. Interestingly, although the growth factor itself was not detected without phosphopeptide enrichment at the whole protein level, ten phosphopeptides matching to different locations within *E. electricus* FGF13 were identified in this study (Table 5.S10) highlighting a possible regulatory mechanism for the highly abundant voltage-gated Na^{+} channel in EO via the C-terminus of the protein. Indeed, mutations in the C-terminal domain of SCN4a have implications in human disease [15].

Teleost fishes have a duplication of the SCN4a gene compared to mammals, namely SCN4aa and SCN4ab. In *E. electricus*, other Gymnotiformes and in mormyroids (which is thought to have evolved EOs independently from Gymnotiformes) SCN4aa is highly expressed in EO and has experienced a higher rate of molecular evolution compared to SCN4ab, which is highly expressed in both muscle and EO, suggesting this Na^{+} channel duplication was important in the independent evolution of electric organs [16]. Using protein alignments among *E. electricus,* other Gymnotiformes, mormyroid, other teleost species and humans, I also observed the C-terminal domain of SCN4aa is highly variable among species (Figure 5.S3). Within the C-terminal domain of SCN4aa, I identified a total of five phosphosites in the *E. electricus* protein, one of these has been previously described in mammals [12], and four are novel. Of these novel phosphosites, two are at residues that are at non-phosphorylatable residues in all other species in the alignment, which could indicate unique functions in *E. electricus*. One phosphosite was localized to a residue that in a subset

of other species is phosphorylatable, including all electric fish in the alignment (*Eigenmannia virescens*, *Sternopygous macrurus*, and *Brienomyrus brachyistius*), which may represent an important function in electric fishes. Finally, I identified one phosphosite in *E. electricus* that is at a phosphorylatable residue in the protein alignment only in *E. virescens* (is a phosphomimetic D in *S. macrurus*). Whether these novel phosphorylation targets are functional adaptions in *E. electricus* and other electric fishes and whether SCN4aa is co-regulated by interactions with calmodulin and FGF13, will be important for future studies.

Potassium channels show unique patterns of protein abundance across EOs

Diversity in the abundance and types of ion channels present in electrocytes has been suggested to be a driver for signal diversity in Gymnotiformes [17], and by extension, speciation. Therefore, it follows that since *E. electricus* has three distinct electric organs that serve distinct functions, perhaps the same underlying principles apply within a species as between. In this study I first examined proteins known to be important for propagating the EOD, and identified several ion channel and ion pump proteins that followed a general pattern of being most abundant in main EO, least abundant in Sachs' EO, and of moderate abundance for Hunter's EO including several Na+/K+-ATPase subunits (e.g. ATP1a2a), voltage-gated Na+ channel subunits (eg. SCN4aa), acetylcholine receptor subunits (e.g. CHRNA1) and acetylcholinesterase (ACHE), and potassium channels (eg. KCNJ12) (Figure 5.1C). This pattern has been previously described for the α subunit of the Na^+/K^+-ATPase, ATP1a2a, in *E. electricus* EOs, and further, it was demonstrated that these enzymes isolated from the different EOs had different kinetic properties in each tissue [18], but outside of ATP1a2a, this pattern has not been demonstrated previously. Multiple α and β subunit

isoforms are expressed in the electric organs, the most abundant of which were ATP1a2a and ATP1b1a (see Figure 5.1C, Table 5.S4 and Table 5.S7). Although the magnitude of change in abundance between Hunter's EO and Sachs' EO was not as great as main EO to Hunter's/Sachs', most ATPase α and β subunits identified shared this pattern of main > Hunter's > Sachs', as did most acetylcholine receptor subunits, acetylcholine esterase, and sodium channel proteins. It is interesting to note that when the voltage-gated sodium channel was first isolated from *E. electricus* EO, only an α-subunit was detected (no β-subunit)[11]. With these analyses I was able to detect the presence of two β-subunits (Figure 5.1C) SCN4ba and SCN1B, both mid-range in terms of total abundance (Table 5.S6). It is interesting to note that previous studies have localized two isoforms of the Na+/K+-ATPase α-subunit present in *E. electricus* electrocytes, one isoform predominately present on the innervated membrane face (an ATP1a1-type), and the other being highly abundant on the non-innervated membrane face (ATP1a2a) [19], although the function of the ATPase at the innervated membrane face have yet to be elucidated. Additionally, I was able to quantify additional protein isoforms present in the three EOs and muscle (a total of five α-subunits and three β-subunits) (Figure 5.1C), albeit, they are less abundant than ATP1a2a (Table 5.S6), a finding that likely reflects the sensitivity of an unbiased, quantitative "shotgun" proteomics approach. Further studies, including localization and kinetic studies, will be useful to elucidate function of these distinct isoforms.

Interestingly, the general pattern of electrogenic proteins being most abundant in main EO and least abundant in Sachs' EO is broken when I examined potassium channels (Figure 5.1C). Three potassium channel proteins were observed, notably an inwardly-

rectifying potassium channel (KCNJ12), and two voltage-gated K+ channels (KCNA1 and KCNH6), KCNA1 and KCNJ12 is most abundant in main EO while KCNH6 is most abundant in Sachs'. This is interesting because one patch clamp study of the innervated membrane of electrocytes revealed a small, low density, voltage-gated outward K^+ currents in a subset of patches examined [20], despite earlier studies failing to detect such a current [10, 21]. Expression of an mRNA transcript for a $K_v1.1$ (KCNA1) in EO was confirmed later, and demonstrated to encode a functional protein when expressed in a heterologous system [22]. Although I have evidence that these proteins are indeed translated and are of reasonable abundance (and is greater total abundance than KCNJ12 in some of the EOs, Figure 5.S4), the function of these proteins in EOs remains elusive. It has been speculated that the channels may be under continuous inhibition that is relieved under certain physiological conditions [22]. To take this speculation further, assuming the translated protein is properly inserted into the plasma membrane, the potassium channels may be under circadian control as other channels in Gymnotiformes. Sodium channels in *Sternopygous* are only trafficked to the membrane during periods of activity [23] and as Gymnotiformes are nocturnal, it is plausible to speculate that these potassium channels are not in the membrane or are deactivated when everyone is doing their experiments during the day. Regardless, the discovery of two voltage-gated potassium channels being translated in high abundance in EOs without a known physiological role of this channel type is evidence that this is an area that needs to be revisited.

From an ion transport standpoint, the Sachs' EO seems adapted to reduce energetic costs due to its relatively high frequency of action potential compared to main EO, as a massive influx of sodium is costly in the form of ATP hydrolysis by the Na^+/K^+-ATPase, the

Sachs' EO reduces energetic demands by reducing the influx of sodium by having fewer sodium channels (and thus requires fewer Na^+/K^+-ATPase). On the other hand, the main EO appears to be less energetically conservative, being highly abundant in sodium channels (AChR and voltage-gated), and Na^+/K^+-ATPase. Taken together, these differences in protein abundance of ion channels and pumps across the three EOs may reflect different demands on the cells found in each EO, and a trade off between being continuously generating low voltage while minimizing energetic costs and intermittently generating high voltage at a higher energetic cost.

Global comparison of the proteome and phosphoproteome reveal a few hundred non-electrogenic proteins that may play roles in electric organ development and/or function.

Very little is known about the molecular differences among the three distinct EOs in *E. electricus.* Efforts in microscopy revealed that electrocytes from the main and Hunter's EOs were packed tightly to one another while electrocytes from the Sachs' EO were more separated from one another with abundant extracellular matrix [21], which may contribute to the differences in voltage generated in the weak and strong voltage organs. Additionally, electrocytes in the Sachs' EO were larger in size than those of the main EO and had larger invaginations [21]. Further, action potentials from main and Sachs' EO were reported to be similar, but action potentials occur much more rapidly in main EO compared to Sachs'—the mean action potential duration from main EO electrocytes were 1.5 msec and in Sachs' EO electrocytes 2.2 msec [11, 21]. I hypothesized that given the and demands on the strong versus weak EOs is quite different (intermittent, strong voltage versus continuous, weak

voltage), that there are underlying differences in protein abundance and protein phosphorylation that contribute to these distinct functional roles in *E. electricus.*

To tease apart these differences I used a variety of clustering approaches. First I clustered all quantified proteins and phosphopeptides in both fish (Figure 5.2A and 5.2B, respectively) to see how the EOs and muscle relate broadly to one another, and found that muscle is most distinct from all tissues, and that among the three EOs, Hunter's and Sachs' cluster most closely together with main EO being most distinct. I then considered the degree overlap among the top 10% most abundant quantified proteins and peptides in both fish (Figure 5.2C and 5.2D). I observed both at a protein and phosphopeptide level that Hunter's EO was the least unique of the four tissues tested, as it had the fewest unshared entries. At a protein-level, in both fish (Figure 5.2C), I observed the largest shared group in the Venn diagram were proteins shared among all three EOs and muscle. Interestingly, at a phosphopeptide level in both fish (Figure 5.2D), I saw distinct tissue-specificity, with the largest shared group in the Venn diagram belonging to phosphopeptides shared among the three EOs only and not muscle, which indicates that there is distinct protein phosphorylation occurring in EOs compared to muscle.

I used k-means clustering to bin proteins and phosphopeptides that had at least two-fold difference in abundance between the highest-expressing and lowest-expressing EO (Figure 5.2E and 5.2F). I observed clusters of unique proteins and phosphopeptides highly expressed in either main EO, Sachs' EO or muscle/Hunter's EO emerge, illustrating that there are differences in protein abundance and phosphorylation among the three EOs. I focused on these bins of differentially abundant proteins and phosphopeptides, notably the

main EO and Sachs' EO bins, which represent groups of proteins important for intermittent, high-voltage (main EO) or continuous, low-voltage (Sachs' EO) discharges.

Proteins and phosphopeptides highly abundant in main EO

I identified several proteins important for EOD in the main EO protein cluster (Figure 5.2E, cluster 2), including voltage-gated Na^+ channel alpha and beta subunits (SCN4aa, SCN4ab, and SCN1B), acetylcholine receptor subunits (CHRNA1), acetylcholine esterase, and the abundant beta subunit for the Na^+/K^+-ATPase (ATP1b2b). In addition, I observed calmodulin to be in the main EO-specific protein cluster, which has been previously described to be incredibly abundant in the EO from *E. electricus*, but differential abundance across EOs has not been shown. I also found several EOD-related proteins in the main EO phosphopeptide clusters (Figure 5.2F, cluster 3), including acetylcholine esterase, acetylcholine receptor subunits (CHRNB1 and CHRND), ATP1a2a, the voltage-gated Na+ channel (SCN4aa and SCN4ab) and also a chloride channel (CLCN1a) that was not observed in the proteome analysis.

In addition to proteins involved with EOD, I also made some other interesting distinctions with the main EO. First, in main EO-specific protein cluster (Figure 5.2E, cluster 2), I found an α-AMP-activated protein kinase (AMPK) subunit (PRKAA2, a catalytic subunit of AMPK); AMPK is the major energy sensor in cells [24]. In addition, I found protein phosphatase 1aa (PPM1aa), which can dephosphorylate the α-subunit of AMPK [25], and falls in several other signaling pathways. I also identified several other proteins involved in other major signaling pathways, in particular MAPK and NFKB signaling pathways, which overlap significantly. For example, in the main EO phosphopeptide cluster

(Figure 5.2F, cluster 3), I identified phosphorylated peptides for proteins in the MAP-kinase cascade, including for MAP kinase proteins (MAP4K4, MAP3K5, and MAP2K2b), and microtubule-associated protein tau (MAPTA and MAPTB). At a protein level, NFKB2 is highly abundant protein in main EO relative to Sachs' EO, but was not included in this initial clustering as it was not detected in one of the muscle biological replicates.

Additionally, I identified a number of transcription factors in the main EO protein clusters, specifically, CBX6b, a component of the polycomb repressive complex. The polycomb repressive complex 1 has been demonstrated to be involved with maintaining transcriptional repression by chromatin/epigenetic mechanisms [26]. I also identified ZNF644b (a predicted zinc finger transcription factor), mutations in which are associated with myopia in humans [27]. In the main EO phosphopeptide cluster, I identified the transcription factor grainyhead-like 2a (GRHL2a), which is a paralogue of UBP1, which is abundant in Sachs' EO (see below). Grainyhead family transcription factors have important and diverse roles in development, notably in the regulation of genes important for cell junction formation (e.g. tight junctions, and others) [28], which are vital to the development or establishment of polarization in cells [29], of which, the very nature of electrocytes in generating voltages requires cells to be very polarized. I also found an aquaporin protein (AQP8a.2) the main EO proteome cluster, which may reflect my underlying hypothesis, that main EO is less energetically conservative compared to Sachs' EO, and must deal with the repercussions of massive ion flux (and osmotic flux) per action potential.

Finally, in the main EO phosphopeptide cluster, I identified four phosphopeptides for NDRG4. The function of this protein or its phosphorylation is not well understood, but

in humans, NDRG4 mutations are associated with elongated QT intervals and other heart defects [30], [31, 32]—the QT interval is a common metric used to measure myocardial repolarization. In zebrafish, knocking down expression of NDRG4 give rise to fish with heart defects, including reduced heart rate, which is interesting, as zebrafish respond to QT-prolonging drugs with decreased heart rates [30]. But because NDRG4 has been demonstrated to be important for regulating rate of repolarization of cardiomyocytes, perhaps it is important for regulating repolarization of electrocytes (as previously mentioned, one known differences among main and Sachs' EO is that action potentials in main EO are shorter [21]). In total for NDRG4, I quantified 17 phosphorylated peptides, and seven phosphosites (from 10 phosphopeptides) that have not been described in mammals (Table 5.S11).

Proteins and phosphopeptides abundant in Sachs' EO

We also found several interesting proteins in the Sachs' EO-specific protein and phosphopeptide clusters (Figure 5.2E cluster 1 and Figure 5.2F cluster2, respectively). I again found several proteins in numerous signaling pathways abundant in Sachs' EO including ULK1, a serene/threonine protein kinase that (together with AMPK, the catalytic subunit of which is highly abundant in main EO, see previous section) is important for regulating cell autophagy [24]. Although the function of the catalytic subunit of AMPK being highly abundant in main EO, and ULK1 being highly abundant in Sachs' EO, it seems plausible that these differences in proteins important for cellular energetics reflects differences in tissue physiology given their intermittent versus continuous use, respectively. Perhaps the role of ULK1 in autophagy reflects both the difference in

electrocyte size in Sachs' EO vs. Main EO (Sachs' electrocytes being larger) and tissue use (Sachs' EO being used continuously perhaps means proteins need to be replaced more frequently). I also identified a Rho GTPase activating protein (ARHGAP21a) and two guanine nucleotide exchange factors (ARHGEF1b, and ARFGEF1). I also identified the transcription factors STAT3 and UBP1, a grainyhead family transcription factor and a paralogue of GRHL2a (highly abundant in Main EO, see previous section). UBP1 expression has been demonstrated to determine intercalated cell development and polarity in Xenopus, notably in which membrane the H^+-ATPase is localized [33]. Although the underlying mechanism is unknown, perhaps UBP1 in Sachs' EO and GRHL2 in main EO are performing similar but distinct requirement in development or maintenance of cell polarity in each tissue type.

The most abundant phosphopeptides in Sachs' EO (Figure 5.2F, cluster 2) belonged to proteins that are implicated in muscular diseases. These include ANO5b, a protein in which mutations have been demonstrated in muscular dystrophy [34] and dilated cardiomyopathy [35]. Although not detected in the whole proteome experiment, at an RNA level, ANO5b shows increased abundance in Sachs' EO over muscle, main, and Hunter's EO at the RNA level (Table 5.S4) and interestingly its expression in this table indicates its has a similar expression level to cardiac muscle. Also highly abundant in Sachs' relative to other EOs includes CMYA5 (cardiomyopathy-associated 5), which, together with desmin, is thought to coordinate specific kinases and phosphatases in muscle [36].

<u>Phosphopeptides that show differential abundance in EOs compared to protein abundance.</u>

In order to determine whether specific phosphosites showed differential abundance that were independent of changes in protein abundance (defined as a difference between the log2 (tissue/median) phosphopeptide and protein values that was at least 1), I first examined the phosphosites found in proteins known to be important for EOD, i.e., the acetylcholine receptor, acetylcholine esterase, voltage-gated sodium channels, and the Na^+/K^+-ATPase. Generally, the majority of phosphopeptides in these proteins showed similar abundance to that of the protein quantitation (Tables S7 and S10). In the most abundant ATPase (ATP1a2a), three fully-tryptic phosphopeptides (representing three phosphosites) showed differential abundance, one of which appeared increased in abundance in main EO-2 (from biological replicate two, but not one) and two of which appeared reduced in abundance in main EO-2 compared to the protein-level abundance, while remaining unchanged in Sachs' and Hunter's EOs (Table 5.S3). The location of these three phosphosites has been previously described in mammals, however, I was unable to find a description of the function of these three phosphorylation sites in the literature.

In the voltage-gated Na^+ channel protein SCN4aa, I identified four fully-tryptic phosphopeptides (representing four phosphosites) that showed differential abundance compared to the protein abundance, three of these showed reduction in at least one Sachs' EO sample (with no change in other EOs). These distinct phosphosites localized with 100% probability (see search methods), and not previously described in humans [12]; one of these (aVSHAsFLSQIk) falls within the C-terminal domain of SCN4aa. The fourth phosphoprotein showed reduction in main EO-2 (with no change in other EOs), and despite being not fully localized, represents a novel phosphorylated region, as it is not near a known phosphosite in mammals [12].

Finally, I identified two phosphosites in acetylcholine esterase (ACHE), both of which have not been described in mammals, and were fully localized (see methods, and Table 5.S3). One of these phosphosites showed a large decrease in the Hunter's EO-2 sample that was not reflected in its protein concentration, and showed slight decreases in main EO-2 and slight increases in Sachs's EO-2. The second phosphosite showed a decrease in both Sachs' EO (~2 fold). Future studies (perhaps those in Appendex I), will be needed to provide additional biological replicates as well as to shed light on the biological importance of these particular phosphosites.

Desmin phosphorylation shows tissue specific abundance

Differential abundance of desmin protein isoforms and desmin phosphorylation were previously reported in the three EOs of *E. electricus* [37, 38]. I identified two desmin proteins- desmin A and desmin B, which are similarly abundant in all EOs at a whole protein level. When considering phosphopeptides, however, distinct phosphopeptides being more abundant in one tissue vs. another is observed. A total of 70 phosphopeptides for desmin A and eight for desmin B. Of these, there are two phosphopeptides for desmin A that fall into phosphopeptide cluster 1 (high main EO), and two in cluster 4 (high Sachs' EO), and for desmin B, there is one phosphopeptide that fall in cluster 1 (high main EO), and one in cluster 3 (muscle cluster- highest in Hunter's EO, variable in Sachs', and very low in main EO). These data agree with the previous reports showing differential protein and phosphorylation abundance across the three EOs [37, 38], and provides locations of phosphorylated residues in these proteins.

The function of desmin in electrocytes is unknown, as is the function of desmin, generally, especially in fish [36]. Roles for desmin extend beyond structural (desmin is an intermediate filament protein): it is involved in calcium homeostasis [36], is abundant in neuromuscular junctions and at the points of attachment between cardiomyocytes [39], is important for mitochondria structure and function, and can regulate the expression of genes important in myogenic and cardiogenic regulation [40]. Mutations in desmin are associated with human disease, and various post-translational modifications including protein phosphorylation have been implicated in desmin-related disease [40].

Hunter's EO abundant in muscle-specific proteins compared to other EOs

It is interesting to note the k-means clusters (both at the whole proteome and phosphoproteome resolution), Hunter's EO consistently expresses proteins intermediately between main and Sachs' EO and also more similarly to muscle compared to Sachs' and main EO. This was not revealed in the hierarchal clustering analysis (Figure 5.2A and B) in which both Sachs' and Hunter's clustered more closely to muscle than main EO; this is likely because in the hierarchal clustering analysis all quantified proteins were included, whereas within the k-means clustering analysis, only a subset of proteins or phosphopeptides with at least two-fold abundance difference among the EOs, were included. This is interesting in given two factors (1) I used the strong-voltage portion of the Hunter's EO (the section under the main EO) for these analyses, and (2) morphologically, Hunter's EO appears most similar to main EO [21, 41], though no studies have comprehensively characterized or compared EOs. Because earlier reports indicated that Hunter's EO is morphologically more similar to main EO than it is to Sachs' EO, and because

I used the strong-voltage portion of the Hunter's EO for these experiments, I anticipated that main EO and Hunter's EO would be extremely similar.

Correlation of mRNA and protein abundance

Given that recent research efforts have focused on mRNA sequencing to understand the repertoire of genes expressed in electric organs of *E. electricus* [8, 42], I wanted to determine whether the protein abundance values correlated with mRNA expression values (see Supplemental Materials). For this purpose, I compared $\log_2$ (EO/muscle) ratios for both the RNA expression and protein abundance levels of the expressed products from the 2866 gene models for which I had protein abundance data. My results (Figure 5.S5 A and B) show that the majority of gene models are not differentially expressed at either an mRNA or protein level. Similar numbers of gene models are differentially expressed only at the protein level, only at the RNA level, or at both the RNA and protein level, in any given EO. The fewest, and perhaps most intriguing, gene models are differentially expressed at both the RNA and protein levels, but in opposite directions.

For example, in main EO, I saw a putative FXYD protein (gene model scaffold112.g65) display discordant RNA and protein abundance values relative to muscle. The annotation of this particular protein is a challenge because not only is it short (my gene model predicts 131 amino acids, which is longer than the zebrafish protein of 90 amino acids), blast searching of zebrafish FXYD proteins reveals these short proteins are highly variable across species. FXYD proteins, also known as the gamma subunit of the Na^+/K^+-ATPase, are very small proteins that co-purify with the Na^+/K^+-ATPase, but for which a clear function has not yet been elucidated, however, they do have a modulatory function of

Na^+/K^+-ATPase [43]. At the mRNA level in main EO, this gene model appears downregulated compared to muscle. However, at the protein level it appears upregulated. A similar result was observed for Hunter's EO, but the results were less dramatic for that of main EO, and in Sachs' EO, differential abundance at the protein level (but not mRNA level) was observed. In the quantitative proteomics experiment, this FXYD protein is highly abundant in main EO compared to Sachs' or Hunter's only in one of the biological replicates (Table 5.S6), and is of similar abundance across all three EOs in the other biological replicate. Further effort in confirming the identity of this protein will need to be made, however, given the FXYD protein family is comprised of sequences that are short and highly variable, this will be a challenge; the best blast match to this protein is FXYD6. However it is important to note that at an mRNA level, this gene model is not lowly expressed- it's abundance is extremely high in muscle, and quite high in the three EOs. I also see that in main EO, ATP1a3a (alpha subunit of Na^+/K^+-ATPase) the protein and RNA abundance ratios are discordant (highly abundant in main/muscle at a protein level, and lowly abundant at an RNA level), which does not appear differentially expressed in either protein or RNA in Sachs' or Hunter's EOs.

Interestingly, I also saw several of the previously discussed transcription factors with discordant protein and mRNA abundances. For example, the transcription factor UBP1 is differentially expressed at the protein level in Sachs' EO, but not in at the mRNA level. In addition, UBP1 is classified as not differentially expressed at the mRNA or protein level in both main and Hunter's EO (or appears slightly downregulated in main and Hunter's). Further, the transcription factor ZNF644b displays low mRNA but very high protein abundances in the main EO, and a similar pattern for one Hunter's EO bioreplicate. If future

experiments with more biological replicates validate the discordance in mRNA and protein levels, this may represent post-transcriptional regulation of these transcription factors, Na^+/K^+-ATPase alpha subunits, and Na^+/K^+-ATPase gamma subunits that may be important for their distinct functions (see Chapter 6, Future Directions).

These data represent the first quantitative proteomic and phosphoproteomic description and comparison of electric organs in any electric fish species. The findings in this study suggest key differences among the three EOs in *E. electricus*. The Sachs' EO appears to be more energetically conservative than the strong-voltage main EO, by expressing fewer Na+ channels, and thus, requiring less ATP hydrolysis by having fewer Na^+/K^+-ATPases to maintain resting membrane potential. This finding is consistent with the key functional differences between Sachs' and main EOs, Sachs' being used continuously for navigation and communication, while main EO is used only intermittently for strong-voltage discharges in predation and defense. I also find differences among the Sachs' and main EOs that reflect differences in signal transduction, as well as transcription factors that may be important in tissue form and function, given their known roles in establishing cell polarity and cell-cell contacts (UBP1 and GRHL1, respectively). Curiously, I identified interesting patterns of potassium channel expression relating to two voltage-gated potassium channels, one of which is expressed most abundantly in Sachs' EO; this finding is curious given that voltage-gated potassium current has been shown to be a very minor current in Sachs' EO. The story surrounding the voltage-gated K^+ channels I detected in this study will be an interesting area of further study- to elucidate if these channels function in electrocytes, and why they are so abundant. Finally, this study describes several new phosphorylation sites in proteins important for electric organ discharge. This study,

together with other recent studies offering molecular characterization of electric organs from a variety of species (Chapters 3 and 4), gives the field of electric fish biology a strong finhold in the 'omics' era.

Acknowledgements

This project has been funded by NSF Grant MCB No. 1144012 (MRS), the Morgridge Graduate Fellowship (LLT) and the University of Wisconsin Genetics NIH Graduate Training Grant (LLT). I graciously would like to thank the UW Mass Spectrometry Facility for aid in study design and implementation, and the UW Sequencing Facility efforts in generating the long-range paired-end 454 reads incorporated into the assembly in this study.

SUPPLEMENTAL MATERIALS AND METHODS:

Contents:

1. Genome Assembly- improvement, and gene prediction

1A: Genome Scaffolding and Gap Filling

In order to improve on the existing *E. electricus* genome assembly in an effort to then improve downstream gene model prediction, additional raw sequencing reads were incorporated into the publically available genome assembly (termed "SOAPdenovo2 assembly). Scaffolding of the SOAPdenovo2 assembly was performed using SSPACE (Standard v3.0) [44], incorporating both Illumina 2.5 kb insert mate pair libraries and 454 libraries, with the command -z 500 -g 3 -v 1 -T 27 and following library information given for–b:

```
Lib1 bowtie run196.eel-MP-1_NoIndex_L002_R1.fastq run196.eel-MP-1_NoIndex_L002_R2.fastq 2500 .2 RF
Lib2 bowtie run196.eel-MP-2_NoIndex_L003_R1.fastq run196.eel-MP-2_NoIndex_L003_R2.fastq 2500 .2 RF
Lib3 bowtie run196.eel-MP3_NoIndex_L004_R1.fastq run196.eel-MP3_NoIndex_L004_R2.fastq 2500 .2 RF
Lib4 bowtie 1.454Reads.qual_len.fwd.fastq 1.454Reads.qual_len.rev.fastq 12000 .333 RR
Lib5 bowtie 2.454Reads.qual_len.fwd.fastq 2.454Reads.qual_len.rev.fastq 12000 .333  RR
Lib6 bowtie 1.TCA.454Reads.qual_len.fwd.fastq 1.TCA.454Reads.qual_len.rev.fastq 12000 .333 RR
Lib7 bowtie 2.TCA.454Reads.qual_len.fwd.fastq 2.TCA.454Reads.qual_len.rev.fastq 12000 .333 RR
```

Following scaffolding, gap closing was performed with GapCloser (v1.12) [45] using both short (2x100) Illumina paired-end reads and Illumina MP reads (2.5 kb insert) with the following configuration file:

```
# cutoff of pair number for a reliable connection (at least 3 for short insert size)
pair_num_cutoff=5
#minimum aligned length to contigs for a reliable read location (at least 32 for short insert
size)
map_len=35
#a pair of fastq file, read 1 file should always be followed by read 2 file
q1=/home/ltraeger/raw_data/run196.eel-MP-1_NoIndex_L002_R1.fastq
q2=/home/ltraeger/raw_data/run196.eel-MP-1_NoIndex_L002_R2.fastq
#another pair of fastq file, read 1 file should always be followed by read 2 file
q1=/home/ltraeger/raw_data/run196.eel-MP-2_NoIndex_L003_R1.fastq
```

```
q2=/home/ltraeger/raw_data/run196.eel-MP-2_NoIndex_L003_R2.fastq
#another pair of fastq file, read 1 file should always be followed by read 2 fi$
q1=/home/ltraeger/raw_data/run196.eel-MP3_NoIndex_L004_R1.fastq
q2=/home/ltraeger/raw_data/run196.eel-MP3_NoIndex_L004_R2.fastq
[LIB]
#average insert size
avg_ins=230
#if sequence needs to be reversed
reverse_seq=0
#in which part(s) the reads are used
asm_flags=4
#use only first 100 bps of each read
rd_len_cutoff=80
#in which order the reads are used while scaffolding
rank=1
# cutoff of pair number for a reliable connection (at least 3 for short insert size)
pair_num_cutoff=3
#minimum aligned length to contigs for a reliable read location (at least 32 for short insert
size)
map_len=32
#a pair of fastq file, read 1 file should always be followed by read 2 file
q1=/home/ltraeger/raw_data/run141.eel2.s_1_1_sequence.fastq
q2=/home/ltraeger/raw_data/run141.eel2.s_1_2_sequence.fastq
#a pair of fastq file, read 1 file should always be followed by read 2 file
q1=/home/ltraeger/raw_data/run141.eel2.s_2_1_sequence.fastq
q2=/home/ltraeger/raw_data/run141.eel2.s_2_2_sequence.fastq
#a pair of fastq file, read 1 file should always be followed by read 2 file
q1=/home/ltraeger/raw_data/run141.eel2.s_3_1_sequence.fastq
q2=/home/ltraeger/raw_data/run141.eel2.s_3_2_sequence.fastq
```

This scaffolded and gap-closed assembly will be referred to later in this text as “GapClosed Assembly”.

1B: Gene Prediction and Annotation

For the purpose of gene prediction, RNA sequencing reads from eight tissues of *E. electricus* (brain, spinal cord, heart, skeletal muscle, main EO, Sachs' EO, Hunter's EO, and kidney, see two papers) were mapped to the scaffolded and gap-filled genome using STAR (v2.4.0) [46], using default parameters. Alignments from all eight tissues were merged, and extrinsic hints for "intron" and "exonpart" features were generated using custom scripts from the STAR output "SJ.out.tab" files and "Aligned.out.sam" files, respectively.

In addition to transcript hints, I also incorporated protein evidence based on Human proteins (Ensembl release GRch37.73), zebrafish proteins (Ensembl release Zv9.73) , and all other Gymnotiform protein sequences available at NCBI Taxonomy Browser (downloaded Feb. 7, 2014). To do this, I followed the tutorial offered by the AUGUSTUS makers:
http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=Augustus.IncorporateProteins

Briefly, all proteins sequences were blasted against the genome assembly (tblastn -evalue 1e-15 –threshold 999 –max_intron_length 3000 –max_target_seqs 10). Top hits were parsed, and then Exonerate (v2.2.0)[47] was run on the top hits using default parameters. The output from exonerate was converted to hints for AUGUSTUS. Protein hints were generated with scripts from Katharina Hoff (University of Greifswald), personal communication.

AUGUSTUS (v2.6) [48] was run with the following parameters: "--species=human --codingseq=on --gff3=on --alternatives-from-evidence=true --allow_hinted_splicesites=atac

--UTR=on --uniqueGeneId=on," with the following parameter setting in the extrinsic cfg file:

```
start       1   1       M  1    1e+100  RM  1   1     E 1     1     W 1     1     P  1     1e3
stop        1   1       M  1    1e+100  RM  1   1     E 1     1     W 1     1     P  1     1e3
tss         1   1       M  1    1e+100  RM  1   1     E 1     1     W 1     1     P  1     1
tts         1   1       M  1    1e+100  RM  1   1     E 1     1     W 1     1     P  1     1
ass         1   1       M  1    1e+100  RM  1   1     E 1     1     W 1     1     P  1     100
dss         1   1       M  1    1e+100  RM  1   1     E 1     1     W 1     1     P  1     100
exonpart          1   .997    M  1    1e+100  RM  1   1     E 1     1e2     W 1     1.007 P  1   1
exon        1   1       M  1    1e+100  RM  1   1     E 1     1e4     W 1     1     P  1     1e4
intronpart   1 1       M  1    1e+100  RM  1   1     E 1     1     W 1     1     P  1     1
intron      1   .3      M  1    1e+100  RM  1   1     E 1     1e6     W 1     1     P  1     100
CDSpart     1   0.985   M  1    1e+100  RM  1   1     E 1     1     W 1     1     P  1     1e5
CDS         1   1       M  1    1e+100  RM  1   1     E 1     1     W 1     1     P  1     1
UTRpart     1   .96     M  1    1e+100  RM  1   1     E 1     1     W 1     1     P  1     1
UTR         1   1       M  1    1e+100  RM  1   1     E 1     1     W 1     1     P  1     1
irpart      1   1       M  1    1e+100  RM  1   1     E 1     1     W 1     1     P  1     1
nonexonpart  1 1       M  1    1e+100  RM  1   1.01     E 1     1     W 1     1     P  1     1
genicpart    1 1       M  1    1e+100  RM  1   1     E 1     1     W 1     1     P  1     1
```

Annotation

Gene names were assigned to the AUGUSTUS gene models by comparison to *D. rerio* ENSEMBL Zv9 build 79 protein sequences, in a method previously described [8].

1C: Analysis of improved genome, gene models

Mass spectrometry approaches rely on a reference proteome to match experimentally derived spectra to a corresponding protein sequence. To aid in maximally

identify mass spectra, I aimed to improve upon the existing *E. electricus* genome assembly (termed "SOAPdenovo assembly", for the purposes of this manuscript)[8], by incorporating additional sequencing reads to further scaffold followed by a round of gap filling (termed "gap filled assembly", see methods). In order to compare the SOAPdenovo2 genome assembly to the assembly obtained after a round of scaffolding and gap filling, I used Quast (v2.3) [49] to generate statistics about the SOAPdenovo2 assembly, the assembly after scaffolding only, and the assembly post scaffolding and gap-filing (Table 5.S1,A). Compared to the previous SOAPdenovo assembly, the gap-filled assembly showed significant improvement, having significantly increased contig and scaffold lengths (longest contig length ~7 times longer and longest scaffold length ~3.5 times longer in gap-filled assembly compared to SOAPdenovo assembly) and a reduced number of contigs and scaffolds (~3.5 times and 2.5 times fewer, respectively); in other words, the new gap-filled assembly more of the assembly in longer assembled pieces.

In addition, I wanted to determine whether the AUGUSTUS gene models predicted in this study were an improvement over the previously published gene models. To that end, predicted proteomes from each source were blasted against *D. rerio* proteins (ENSEMBL Zv9 build 79), and compared to one another based on blast hit scores (Table 5.S1,B). I used blastp with default parameters, and extracted only top hits for every *E. electricus* protein along with their corresponding score. Then, using custom scripts, I compared the top score for each *D. rerio* protein ID. In this approach, I found that 6714 times the "old" and "new" gene models had identical blastp scores, 7365 times the "new" gene models had a better score, and 2619 times the "old" gene models had a better blast score. In other words, , I found that ~84% of the time, gene models predicted in the gap-filled assembly were as

good or better thantheprevious gene models in the SOAPdenovo assembly.the*E. electricus* proteome database for mass spectral searching was derived from these new gene model predictions.

1D: Calculation of RNA Expression Levels

The GFF output from AUGUSTUS was converted to GTF format using custom scripts for input into HTseq [50]. Expression values were calculated at the gene level (as opposed to transcript level) for every AUGUSTUS gene model using the previously described STAR mapped RNAseq read to genome and the htseq-count command from HTSeq ("-m intersection-strict –a 3 –t exon –s no –i gene_id") (v0.6.1p1)[50]. Normalized for library size with DESeq (v1.18.0) [51], and further normalized for transcript length (exonic length) to generate "reads per kb transcript". These values are found in Table 5.S4.

2. Protein isolation, and mass spectral analyses

2A: Protein isolation and digestion

Overview of method used depicted in Figure 5.S1. Two ~24 inch specimen were used in this study, according to animal protocol (Animal protocol number M01657). Tissue was dissected, flash frozen in 50mL conical tubes, and stored at -80°C for later use. Protein isolation through protein digestion occurred on the same day as to avoid freeze/thaw on intact proteins, and was performed on main EO, Sachs' EO, Hunter's EO, and skeletal muscle from two fish using a methanol/chloroform extraction method. A small piece of each tissue was removed from the stock tissue, was pulverized in liquid nitrogen using a ceramic

mortar and pestle. Five mL of urea lysis buffer was added (8M urea, 25 mM Tris pH 8, 100 mM sodium chloride, 25 mM sodium fluoride, 10 mM sodium pyrophosphate, 50 mM β -glycerophosphate, and 1-Roche Complete EDTA-free Protease Inhibitor Cocktail Tab and 1-Roche PhosStop Inhibitor tab per 20 mL buffer). Samples were homogenized using a Tissue Tearer for 1-2 minute at 4 C. In a 50 mL conical tube, methanol/chloroform extraction was performed by adding 4 volumes MeOH and 1 volume CHCl3, followed by vortexing, 3 volumes of H2O were added, followed by vortexing and then spinning for 5 minutes at 4000 RCF. After spinning, the top, aqueous phase was removed, leaving the interphase intact. Excess methanol was added followed by vortexing, and spinning for 5 minutes. The supernatant was removed, and the pellet was washed a second time with excess methanol, vortexted, and centrifuged. Pellets were dried slightly, and were resuspended in 8M Urea and 50 mM ammonium bicarbonate containing Roche PhosStop (1 tab/20 mL). Protein concentrations were determined using a BCA assay (Thermo Scientific), and since BCA assays require lower urea concentrations, a small aliquot of the sample in 8M urea was used and diluted down to 1M urea. Post protein quantitation, a total of 1.5 mg was diluted to 1.5 M urea in 50 mM ammonium bicarbonate containing Roche PhosStop, and incubated with dithiothreitol (DTT) at a final concentration of 5 mM for 35 min at 65°C. Alkyation was performed by incubating with iodoacetamide at a final concentration of 12.5 mM, in the dark for 1 hour at room temperature. The reaction was quenched adding DTT to a final concentration of 10 mM. Samples were digested trypsin (Promega) and Lysyl endopeptidase (Wako) with at 1:100 ratio of protein to each enzyme, at 37°C overnight.

Reactions were stopped by acidifying samples with formic acid to a final concentration of 0.5 mM. Samples were desalted using solid phase chromatography (Sep-

Pak 3cc C18 cartridge, Waters), such that 1/3 (500 ug protein) was in one column, and 2/3 (1 mg protein) was in another. 3cc C-18 columns were used for cleanup, and the following protocol was used: 3 volumes 100% acetonitrile (ACN); 2 volumes 75% ACN, 0.05% formic acid (FA); 3 volumes 0.1% trifluroacetic acid (TFA). To each digest, 1 volume 0.1% TFA was added, and then loaded on the column. The first flow through was collected and run through the column a second time. The sample was washed with three volumes of 0.1% TFA, and eluted with 1 mL 75% ACN, 0.05% FA, twice, followed by one elution with 1mL 90% ACN. Samples were dried in a speedvac, and stored at -80°C until use. The 500 ug aliquot was used to quality check samples prior to TMT labeling.

2B: Labeling peptide samples with unique tandem mass tag

The eight 1mg samples were then labeled with a unique tag TMT (ThermoScientific, TMT-10 plex, lot PI202555). From the first eel, muscle #1 (TMT-126), main #1 (TMT-127N), Sachs' #1 (TMT-127C), Hunter's #1 (TMT-128 N), and from the second eel, muscle #2 (TMT-128C), main #2 (TMT-129N), Sachs' #2 (TMT-129C), Hunter's #2 (TMT-130N). In total, 3 tubes of 0.8 mg label were used to label 1 mg input. The reactions were scaled up accordingly, and followed manufacturers protocol with the following exceptions: The dried peptides were brought up in 300 uL 200 mM TEAB. Three tubes of each label (0.8 mg tube) were resoluablized in 41 uL neat ACN, vortexted, and centrifuge to collect. TMT labels were added to each sample, vortexed, and incubated at room temperature for 2.5 hours, on a shaking table. Reactions were quenched by adding 24 uL 5% oxalamine, and incubating for 15 min at room temperature. To test mixing ratios, 5 uL of each sample was pulled from

each sample and combined in a 1:1:1:1:1:1:1:1 ratio with the other samples. The remaining labeled peptides were frozen at -80°C until later use.

For the TMT test mix, samples were desalted using C18 columns (OMIX C18 100uL tips, Agilent Technologies, Part number A57003100), and run on an Orbitrap for ratio testing. Summed reporter ion intensities were used to normalizing, finding first the sum of all reporter ion intensities for all peptides, finding the median summed value for all eight channels, and then using this to determine how to adjust the volume input to ensure equal ratios were added for all eight channels.

2C: Prefractionation using high pH reversed-phase chromatography

Labeled 1mg samples were thawed, combined in equal ratios based on above calculations, and were dried down to ~200 uL volume in a speedvac. Samples were brought up in ~1 volume buffer A for high pH reversed-phase chromatography. Separation was performed on an HPLC (Waters 2795) fitted a Gemini 5u C18 250x100 column (Phenomenex) with a flow rate of 5 mL per minute. Buffer compositions were as follows: Buffer A, 10 mM ammonium formate in H2O, pH 10; Buffer B, 10 mM ammonium formate in 80% ACN, pH 10. The following gradient was used for separation: 0-2 minutes, 100% buffer A; 2-3 minutes, 0-5% buffer B; 3-23 minutes, 5-60% buffer B; 23-25 minutes, 60-100% buffer B. 25-26 minutes 100% buffer B; 26-27 minutes 0-100% buffer A; 27-35 minutes 100% buffer A. In total, 70-30 second (2.5 mL) fractions were collected. These fractions were combined to 1-minute fractions. For whole proteome analysis, 500 uL of each 1-minute fraction was removed, and 4.5 mL was retained for phosphopeptide

enrichment. 500 uL samples were dried in a speedvac; 4.5 mL samples were lyophilized. All samples were stored at -80°C post drying.

2D: Unenriched, whole protome data acquisition

For whole proteome analysis, fractions (see above section) were combined in the following manner: From 1-minute fraction #5 (combined 30 second fraction 9 and 10) thru 1 minute fraction 31, every sixth fraction was combined such that six final samples were obtained. Peptides were analyzed by nanoLC-MS/MS using the Agilent 1100 nanoflow system (Agilent, Palo Alto, CA) connected to a new generation hybrid linear ion trap-orbitrap mass spectrometer (LTQ-Orbitrap Elite™, Thermo Fisher Scientific) equipped with an EASY-Spray™ electrospray source. Chromatography of peptides prior to mass spectral analysis was accomplished using capillary emitter column (PepMap® C18, 3µM, 100Å, 150x0.075mm, Thermo Fisher Scientific) onto which extracted peptides were automatically loaded. NanoHPLC system delivered solvents A: 0.1% (v/v) formic acid , and B: 99.9% (v/v) acetonitrile, 0.1% (v/v) formic acid at 0.50 µL/min to load the peptides (over a 30 minute period) and 0.3µl/min to elute peptides directly into the nano-electrospray with gradual gradient from 3% (v/v) B to 30% (v/v) B over 154 minutes followed by 10 minutes fast gradient from 30% (v/v) B to 50% (v/v) B at which time a 7 minute flash-out from 50-95% (v/v) B took place. As peptides eluted from the HPLC-column/electrospray source survey MS scans were acquired in the Orbitrap with a resolution of 120,000 followed by HCD-type MS2 fragmentation at 30,000 resolving power of 10 most intense peptides detected in the MS1 scan from 400 to 2000 m/z; redundancy was limited by dynamic exclusion. Activation settings for the optimum TMT-based MS2

fragmentation were as follows, default charge: 2, isolation width: 2 m/z, normalized collision energy: 38 and activation time of 0.1ms. Raw data was directly imported into Proteome Discoverer (v1.4.14) where protein identifications and quantitative reporting was generated. (see section 2G for search details)

2E. Phosphopeptide enrichment by titanium dioxide chromatography

For phosphopeptide enrichment, fractions were combined in the following manner. From 1 minute fraction 5 (combined 30 second fractions 9 and 10) thru 1 minute fraction 31, every fourth fraction was combined, such that four final samples were obtained. I found that despite being in a volatile salt solution, at these volumes (a total of ~31 mL total fraction volume per combined sample) these resulting samples remained too salty forthetitanium dioxide chromatography method, so each fraction was desalted on a 3cc C-18 column. The desalted samples were then enriched for phosphopeptides using titanium dioxide chromatography, in a method based on [52], but with the following modifications: titanium dioxide tips were prepared with 1.5 mg titanium dioxide (10um) per column, packed atop a C8 disk. Desalted fractions were dried to completion, and resolubilized in 100 ul loading buffer (1M glycolic acid, 80% ACN, 5% TFA). The column was washed twice with loading buffer, and then the sample was loaded, and washed twice with 100 ul loading buffer, and washed twice more with 100 ul wash buffer (80% ACN, 1% TFA). Enriched phosphopeptides were eluted twice from the titanium dioxide with 50 ul 1% ammonium hydroxide, and eluted from the C8 disk with 30% ACN, 0.1% TFA. The eluate was acidified with 3.5 ul neat formic acid, and dried using a speedvac.

2F: Phosphoproteome Data Acquisition

Phosphopeptides were analyzed on an Orbitrap Fusion fitted with a nano flow liquid chromatography column, as previously described [53].

2G: Database Development, searching, and FDR estimation

For mass spectral searching, an *E. electricus* protein database was derived from AUGUSTUS gene models (see previous section). Because often a single gene model had more than one possible coding sequence predicted as a reflection of alternative splicing predictions, to avoid losing quantitative power and still retain the limitation of only quantifying uniquely-mapping peptides, all proteins sequences predicted for a single gene model were concatenated such that they represented a large, single sequence in the .fasta file. This input .fasta file was used as the protein database for Proteome Discoverer.

For untargeted, whole proteome analysis, searching was performed with the following parameters: Thermo Proteome Discoverer (v1.4.1.14), using the Seaquest HT search engine platform to interrogate the *E. electricus* database (see above) plus common lab contaminants. Cysteine carbamidomethylation and TMT-specific labeling was selected as static modifications whereas methionine oxidation and asparagine/glutamine deamidation were selected as dynamic modifications. Peptide mass tolerances were set at 10 ppm for MS1 and 0.02 Da for MS2. A decoy database of reverse sequences was used for false discovery rate (FDR) estimation. A minimum of two unique peptides per protein were required for untargeted, whole proteome quantitation, with a minimum peptide score (XCorr threshold) of at least 1.0. For protein grouping, only peptide spectral matches with FDR <= 0.05. Delta CN better than 0.15. Coisolation eluting peptides less than 75%.

For TiOX-enriched peptides, searching was performed as above, except, Phosphorylation was included as a dynamic modification, with no minimum peptide per protein required. In addition, in Proteome Discoverer (v1.4.1.14), a phosphoRS search node was included to provide confidence of phosphosite localization.

2H: Normalization

A table containing reporter ion intensities for all peptide matches was exported from Proteome Discoverer and parsed using custom scripts: a maximum co-isolation filter of 75% was applied, and all values were scaled by scan injection time such that the reporter peak intensities were multiplied by scan injection time. For all reporter ion intensities passing a false discovery rate filter of < 0.05, the median reporter ion intensity for all peptides in each channel was used to normalize all channels towards the median, and then applying a correction factor for each channel to all values in the table. Finally, peptides were filtered on whether the peptide was "used" by Proteome Discoverer. These values were then used to determine and generate a table for the summed reporter ion intensities for each protein grouping. A summary of total peptides and protein groups identified in each experiment is summarized in Table 5.S2.

In order to determine how similar identical tissues from different animals were to one another, for each fish, the median summed reporter ion intensity for each protein group was determined, and then the $\log_2$ (tissue/median) value was taken for each fish. These $\log_2$ transformed ratios were then input into R for examination. First, density plots were generated of the $\log_2$ (tissue/median) values to examine their distribution centered around zero (Figure 5.S2). Next, I clustered these values by protein abundance profile,

Euclidean distances were computed and complete-linkage hierarchical clustering was performed. The resulting heatmap (built with heatmap.2 function from gplots package, (v2.17.0) [54] (Figure 5.2A) showsthenormalization is successful, as (1) identical tissues from both fish cluster together and (2) muscle clusters distinctly from electric organ.

2I: Generation of tissue-specific clusters

Unenriched, whole proteome. Because I was interested in examining proteins of differential abundance among the three EOs, the normalized, log2 (tissue/median) values were further filtered such that only proteins with a minimum of 2-fold difference between the highest and lowest abundant electric organ were retained. These were input into R (v3.1.2) for k-means clustering and resulting heatmap generation using the pheatmaps package (v1.0.2) [55], using Euclidean distance measurement and complete hierarchical clustering with k=3.

TiOX enriched, phosphoproteome: Using the normalized, log2 (tissue/median) values generated above, only phosphorylated peptides were retained in the analysis as were only phosphopeptides that showed at least two-fold difference between the highest and lowest expressing protein. An identical clustering method was use here as was used for the unenriched, whole proteome clustering, except k=4 was chosen.

2J: Correlation of RNA expression and protein abundance values

In order to determine how the RNA expression values obtained (see Supplementary Information section 1D and Table 5.S4) compared to that of the protein abundance values obtained in these experiments, I gathered the RNA expression data for muscle, main EO,

Sachs' EO, and Hunter's EO only the gene models whose protein sequences were identified in these experiments (a total of 2866 proteins, once common contaminant hits were filtered out). I then took the log2 (EO/muscle) ratio for both the normalized RNA expression values and the normalized protein abundance values for both fishes (log2 (EO/muscle, starting with values in Table 5.S6). To account for the difference in dynamic range between RNA sequencing and protein abundance data, a gene model was considered differentially expressed (DE) at the RNA level if it had a log2 (EO/muscle) value of greater than 2 or less than -2 (at least four-fold difference); a protein was considered DE if it had a log2 (EO/muscle) ratio of greater than 1 or less than -1 (at least a two-fold difference).

To visualize the correlation of the RNA and protein abundance data, the aforementioned criteria were used to classify each protein in each tissue (and in each fish). Five possibilities were considered for each gene model (protein group): (1) not DE at the RNA or protein level, (2) DE at a protein level, but not at the RNA level (3) DE at an RNA level, but not at the protein level, (4) DE at both RNA and protein levels (5) DE at both RNA and protein levels, but opposite in direction. Using scripts, each gene model/protein was classified into each of these five bins, and a table of z-values (for color specification) was created for input into R for building plots (Figure 5.S5).

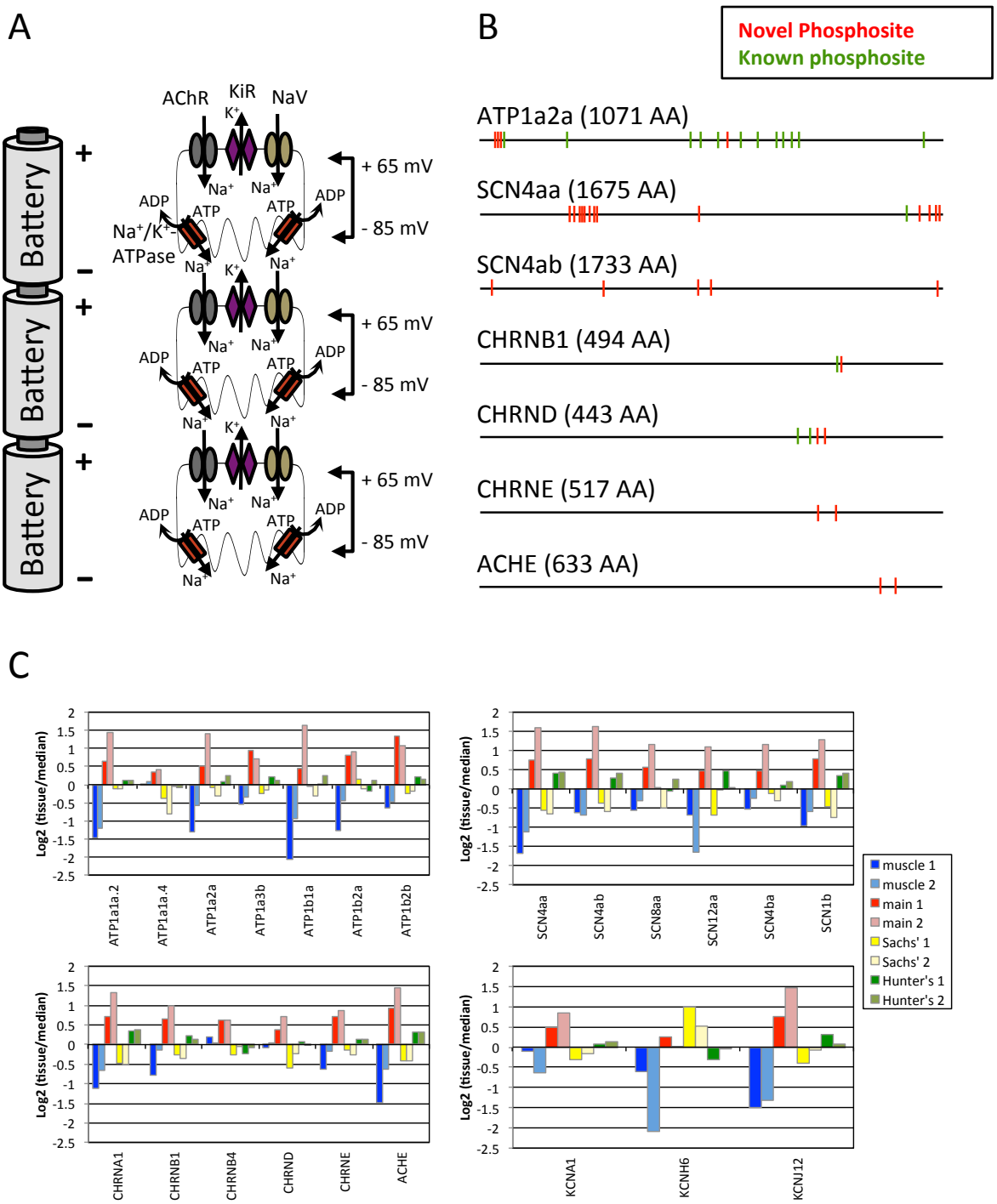


Figure 5.1: Proteins important for electric organ discharge in *E. electricus* have phosphorylated residues not reported in mammals. (**A**). Illustration of electrogenic

proteins in electric organ discharge. Action potential is driven by massive Na^+ influx mediated by Na^+ channels (acetylcholine receptors and voltage-gated Na^+ channels) in the innervated membrane face, while the Na^+/K^+-ATPase work to maintain negative membrane potentials at the non-innervated face. This results in a large transcellular potential difference, and because electrocytes are arranged massively in series, like batteries in a flashlight, voltages summate. (**B**) Relative location of identified phosphorylated residues in electrogenic proteins. Green lines indicate location of phosphorylated residues identified in these data that have been described in mammals [12]. Red lines indicate novel phosphorylated residues in *E. electricus*. Blue indicates phosphosites identified in *E. electricus* that fall within less than five residues of a known phosphosite in mammals. Amino acid length given is for gene model chosen for analysis (see Table 5.S11 for gene model selection, when more than one was possible, if applicable). (**C**) Relative expression for electrogenic ion transporters in EOs and muscle. Generally pattern observed amongst most ion transporters (main > Hunter's > Sachs) is broken by a voltage-gated potassium channel.

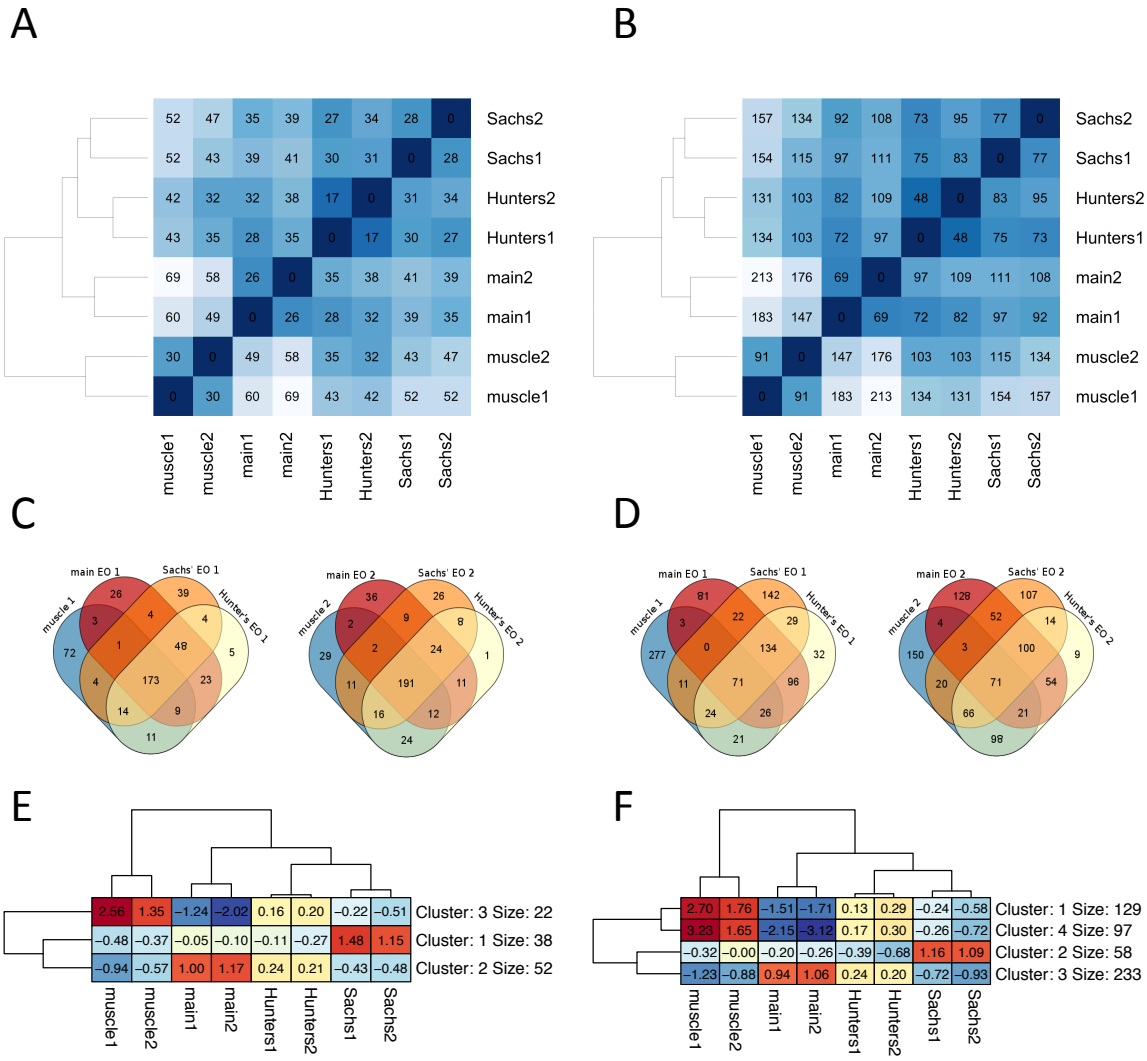
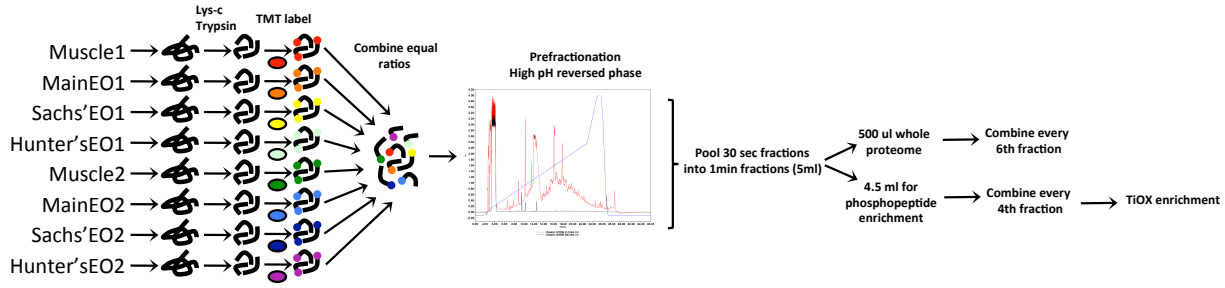


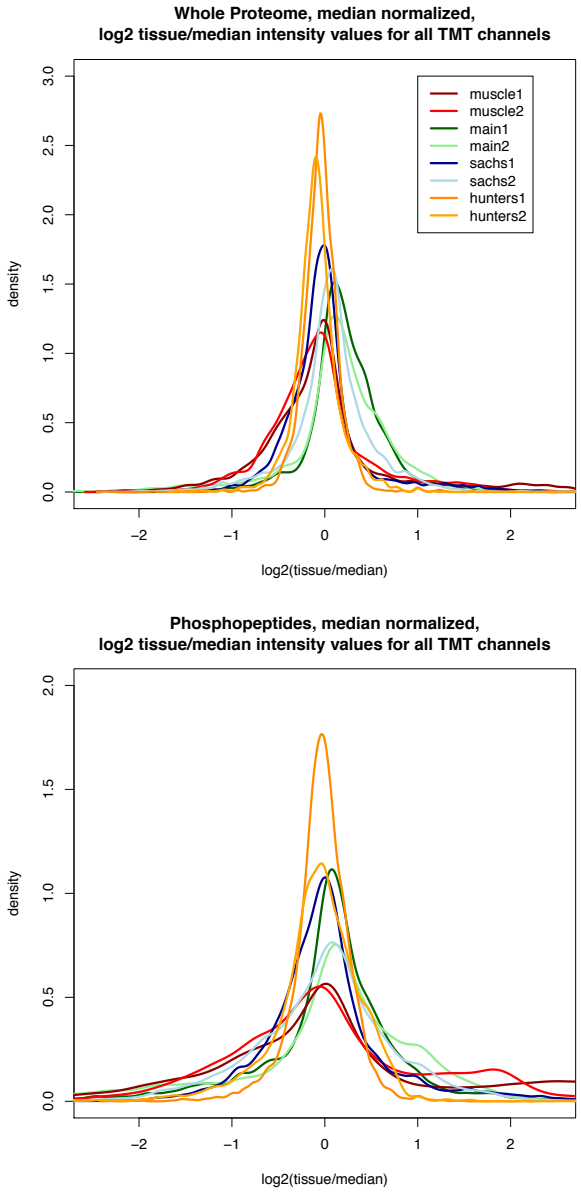
Figure 5.2: Clustering reveals relatedness among three distinct EOs. (**A** and **B**) Clustering of biological replicate tissues by (A) relative protein abundance and (B) relative phosphopeptide abundance**.** Quantitative values for protein groups and phosphopeptides were normalized as indicated in the methods. Clustering was performed using complete-linkage hierarchical clustering, and the numbers inside each square indicate Euclidean distance measurement, darker colors indicating small distances. The data indicates that the identical tissues from two fish cluster together (ie. main EO from eel 1 clusters with main EO from eel 2), and muscle clusters separately from the three electric organs, as expected

in both. Interestingly, the heat map suggests because Hunter's EO and Sachs' EO cluster most closely to one another, that they are more similar to one another and are more distinct from main EO. (**C** and **D**) Venn diagrams showing overlap of top 10% most abundant (C) proteins and (D) phosphopeptides in each tissue and fish. Venn diagrams reveal that Hunter's EO has the least unique proteins and phosphopeptides compared to main, Sachs', and muscle. Protein-level Venn diagrams (C) show that at a protein level, the largest shared group is among proteins most abundant in all four tissues tested (muscle and EO). Phosphopeptide-level Venn diagrams (D) indicate that at a phosphopeptide abundance level, the largest shared group is among the three EOs, indicating there is distinct protein phosphorylation in the EOs relative to muscle. (**E** and **F**). K-means clustering depicts co-regulation in specific EO tissues at a (E) protein-level and (F) phosphopeptide level. Results indicate there are EO-specific patterns of protein and phosphopeptide abundance.

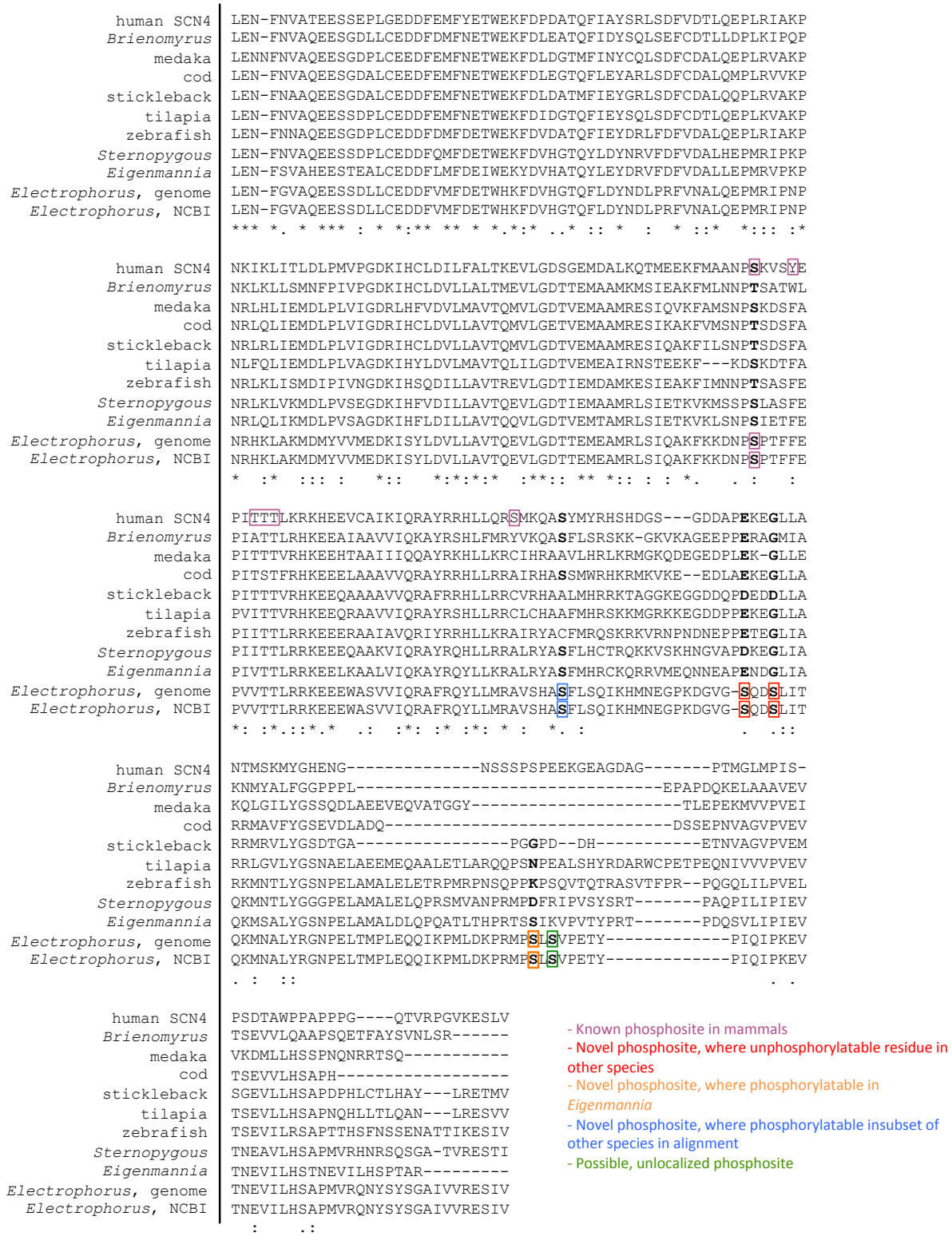


Supplemental Figures:

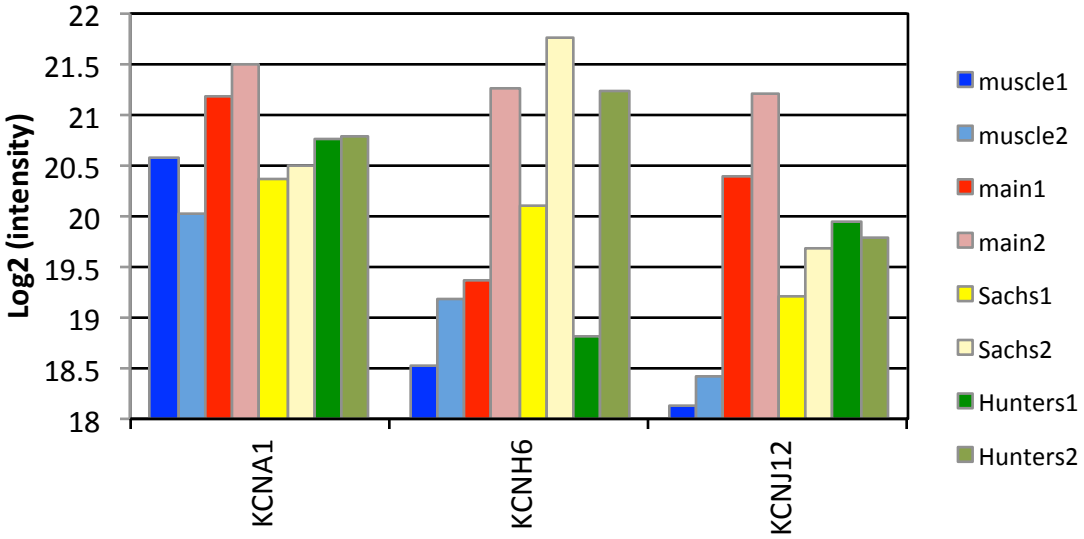
Supplemental Figure 5.S1: Overview of experimental method. Flowchart showing overview of procedures in article.



Supplemental Figure 5.S2: Density plots showing distribution of log2 (tissue/median) normalized intensity values.


human SCN4
Brienomyrus
medaka
cod
stickleback
tilapia
zebrafish
Sternopygous
Eigenmannia
Electrophorus, genome
Electrophorus, NCBI
- Known phosphosite in mammals
- Novel phosphosite, where unphosphorylatable residue in other species
- Novel phosphosite, where phosphorylatable in Eigenmannia
- Novel phosphosite, where phosphorylatable insubset of other species in alignment
- Possible, unlocalized phosphosite

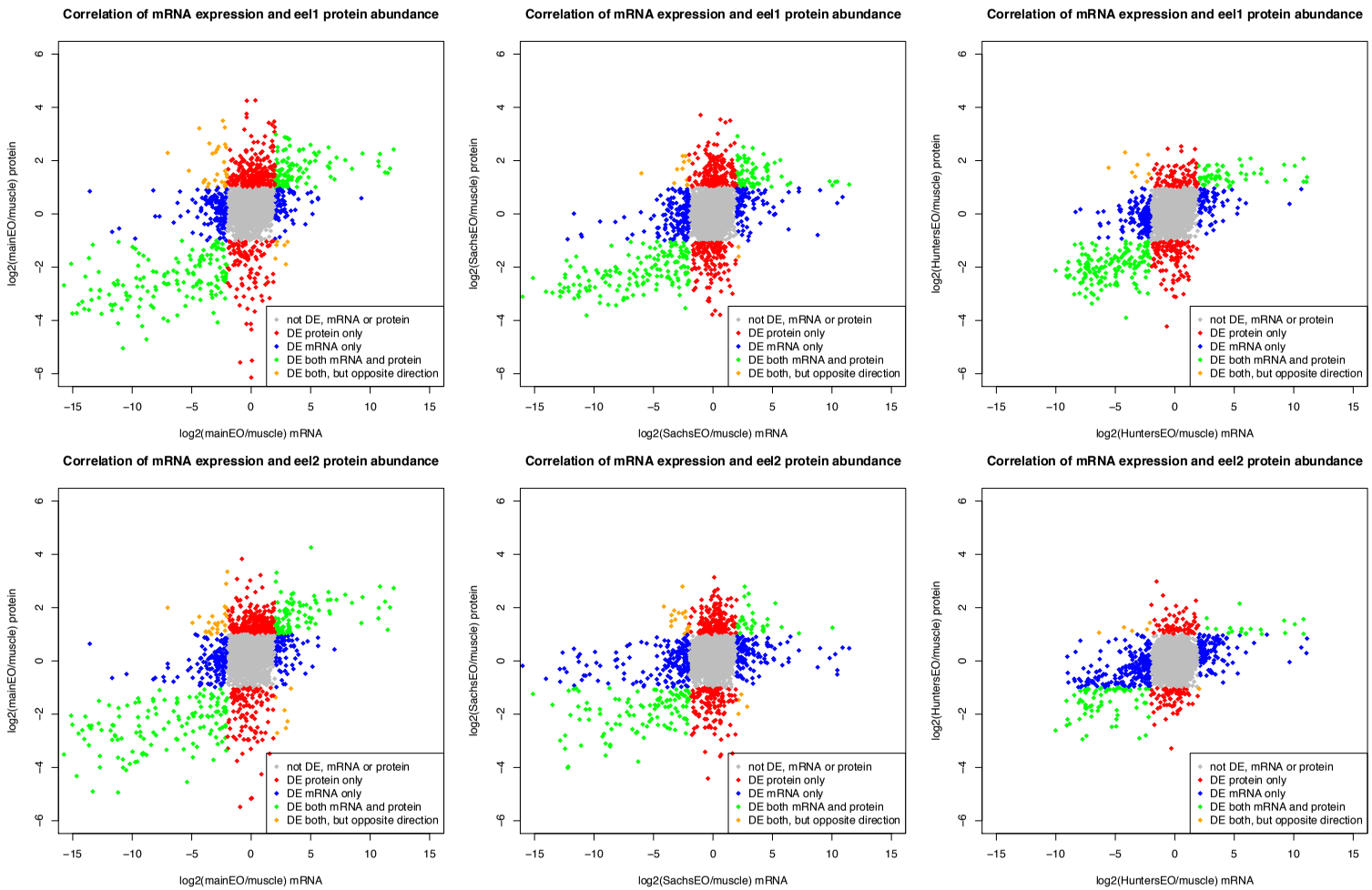
Supplemental Figure 5.S3: Phosphorylation sites in C-terminal domain of *E. electricus* SCN4aa. C-terminal domain of the protein shows variation among species, including among Gymnotiformes (Eigenmannia virescens, and Sternopygous macrurus include). The SCN4aa sequence available at NCBI for *E. electricus* was included for reference. One phosphosite previously described in mammals and four novel phosphosites were localized in the C-terminus of SCN4aa, and are depicted in colored boxes. Red boxes indicated the two localized phosphosites at residues in which other species in the alignment have unphosphorylatable residues. The blue box indicates a localized phosphorylation site at a residue that is also phosphorylatable in a subset of the species included in the alignment (phosphorylatable in the Gymnotiformes the Mormyroid (*Brienomyrus brachyistius*) included in the alignment). The orange box indicates a localized phosphosite at a residue that is phosphorylatable in *E. virescens* (Human shows a phosphorylated residue, however, the alignment is very poor for human). The green box indicates an unlocalized phosphosite possibility. Sequences downloaded from NCBI: Human SCN4a, as ancestral reference (NP_000325.4) and *Sternopygous macrurus* (AAK55442.2). Sequences downloaded from ENSEMBL Build 81: tilapia (ENSONIP00000009933), cod (ENSGMOP00000003144), medaka (ENSORLP00000014568), stickleback (ENSGACP00000004617), and zebrafish (ENSDARP00000134593). Sequences from transcriptome assemblies [8] for and transcript sequences for *Brienomyrus brachyistius* (transcript comp27150_c0_seq1) and *Eigenmannia virescens* (transcript Ev-comp269953_c0_seq20). Transcript assemblies covering a significant portion of the C-terminal end were not identified in the transcriptome assemblies for *Malpterurus electricus* or *B. brachyistius*.



Supplemental Figure 5.S4: Abundance of potassium channels detected in this study.

Values shown are log2 normalized summed reporter ion intensity values in Table 5.S6.

A



B

eel1	main EO	Sachs' EO	Hunter's EO		eel2	main EO	Sachs' EO	Hunter's EO
(1) no DE in RNA or protein	1840	1904	2087		(1) no DE in RNA or protein	1778	1886	2134
(2) DE protein only	399	361	218		(2) DE protein only	440	360	145
(3) DE RNA only	220	253	221		(3) DE RNA only	242	316	346
(4) DE both	255	223	218		(4) DE both	232	145	89
(5) DE both, opposite	36	13	8		(5) DE both, opposite	31	21	6

Supplemental Figure 5.S5: Correlation of mRNA expression and protein abundance. (A). Plots show correlation of mRNA expression and protein abundances for all proteins dectected in unenriched proteomics experiment. See Supplementary Information, section 2J for details on how graphs were generated. (B) Table showing counts of gene models falling into each category depicted on graphs in part A.

Supplemental Tables:

A

SOAP genome only

Assembly	contigs (split on Ns)	scaffolds
# contigs (>= 0 bp)	471054	121323
# contigs (>= 1000 bp)	103972	16715
Largest contig	65235	886512
N50	5954	107900

scaffolded SOAP genome, illumina only

Assembly	contigs (split on Ns)	scaffolds
# contigs (>= 0 bp)	353998	8840
# contigs (>= 1000 bp)	102528	6133
Largest contig	65235	3119439
N50	6126	595548

gapfilled, scaffolded SOAP genome, illumina only

Assembly	contigs (split on Ns)	scaffolds
# contigs (>= 0 bp)	47656	8840
# contigs (>= 1000 bp)	32881	6096
Largest contig	429623	3115075
N50	37312	596078

scaffolded, with illumina and 454

Assembly	contigs (split on Ns)	scaffolds
# contigs (>= 0 bp)	353965	8788
# contigs (>= 1000 bp)	102523	6081
Largest contig	65235	3119439
N50	6127	616204

gapfilled, scaffolded with illumina and 454

Assembly	contigs	scaffolds
# contigs (>= 0 bp)	47607	8788
# contigs (>= 1000 bp)	32849	6044
Largest contig	438728	3115075
N50	37346	613956

B

Gene Model Comparison	Number of Proteins
Old and new gene models 'tie'	6714
New models have better score	7365
Old models have better score	2619

Table 5.S1 Comparison of new genome assembly and gene annotations to the previous assembly. A. Table shows statistics on comparing various assemblies to the original assembly published previously [8]. After incorporating new 454 reads, and a round of scaffolding and gapfilling, this improved assembly has ~3.5 times fewer contigs, 2.5 times fewer scaffolds, the longest contig is ~7 times longer, the longest scaffold is 3.5 times longer. In other words, more of the newest assembly is in longer assembled pieces. Statistics generated by Quast [49]for each genome assembly rendition (see methods): "SOAP genome only" (original genome assembly [8]), "scaffolded SOAP genome, illumina only" (original genome assembly scaffolded with Illumina mate-pair reads only), "gapfilled, scaffolded SOAP genome, illumina only" (scaffolded with illumina plus a round of gap filling), "scaffolded, with illumina and 454" (original genome assembly scaffolded with both Illumina and 454 mate-pair reads), "gapfilled, scaffolded with illumina and 454" (scaffolded with illumina/454 plus a round of gap filling. This is the assembly used in this manuscript) **B.** Comparison of gene models generated in old and new assemblies. Predicted proteins were blasted against zebrafish proteins, and compared to one another based on blast hit store. This comparison revealed that ~84% of the time, the new predicted protein sequences were as good or better than the previous gene model, meaning 84% of the time, they received the same blast hit score to zebrafish, or a better score compared to the previous equivalent gene model.

Unenriched, whole proteome		
	# unique peptides	26668
	# protein groups	2873
TiOX enriched, phosphoproteome		
	# unique peptides	7905
	# unique phosphopeptides	4334
	% enrichment	54.80%
	# unique proteins	2076

Table 5.S2. Peptide and protein group counts for unenriched, whole proteome experiment and TiOX-enriched phosphoproteome experiment.

peptide	protein_group	name	muscle1	muscle2	main1	main2	sachs1	sachs2	hunters1	hunters2	known in mammals	note:
aDFLPGQEsVPILkR	scaffold672.g4	atp1a2a	-2.0935659	-0.7570041	0.04625254	1.45969783	0.9886552	0.18046907	-0.0477846	-0.206316	no?	missed cleavage, highly abundant in one Sachs1
dMTsEFLDDLLR	scaffold672.g4	atp1a2a	NA	NA	0.98578217	2.75372607	-0.5225098	NA	0.38276716	1	yes	tryptic, unoxidized M2. Another peptide with oxidased M2 shows no change. Missing values likely skewing true ratio.
fQLSIHEIEESPtGHILVmk	scaffold672.g4	atp1a2a	-1.8360012	NA	0.04101178	0.39818219	0.38891508	0.26093469	-0.0422118	-0.3187842	yes? (not localized)	tryptic, reduction in one main2.
gVGIISEGNEtVEDIAER	scaffold672.g4	atp1a2a	-1.3646578	0.48849248	0.2063303	0.24964293	-0.2408437	-0.3020711	0.31863618	-1.2973816	yes	tryptic, reduction in one main2.
iLDRcsSIMISGQDIPLNDEWTNAFQR	scaffold672.g4	atp1a2a	NA	NA	0.85648549	2.96006163	-0.0468123	1	0.04534092	NA	no	missed cleavage. Highly abundant in one main2/Sachs2. Missing values likely skewing ratio.
rNsVFQQGMR	scaffold672.g4	atp1a2a	-0.7361037	-1.0669759	0.2441314	0.08786837	0.64967556	0.5906946	-0.29403	-0.0935691	yes	missed cleavage. Reduced in main2, highly abundant in both Sachs1 and 2.
nHMDsGEEPDSEMRk	scaffold610.g4	scn4ab	-1.993129	-3.0234305	0.40575786	1.61639202	-0.2042191	-1.5921731	0.17886329	0.73840339	no	missed cleavage. no change in main EOs. Reduction in Sachs2
nHMDsGEEPDSEmRk	scaffold610.g4	scn4ab	-0.4165437	NA	-0.5706694	1.07302828	0.74611067	-0.8465547	0.32283103	0.52995962	no	missed cleavage. reduction in main1, reduction in sachs2
nHmDsGEEPDSEMRk	scaffold610.g4	scn4ab	-1.6071598	NA	0.46633373	1.669871	-0.3083864	-1.0041401	0.25393548	0.58633991	no	missed cleavage no change
qGRcsFGR	scaffold610.g4	scn4ab	-3.691639	-2.3678562	0.45504889	2.08028348	-0.6684765	-1.5152748	0.73016	0.72261481	no	missed cleavage. Nearly 2fold reduced in sachs2
aVSHAsFLSQIk	scaffold655.g8	scn4aa	NA	NA	0.13393615	1.47442815	-0.1476549	-1.2489777	0.39765472	0.65924301	no	reduction in Sachs2
dGVGSQDsLITQk	scaffold655.g8	scn4aa	-3.9894298	-2.0946122	0.84342927	1.6701729	-1.247117	-1.7100949	0.65874686	0.76074047	no	reduction in both Sachs, >2fold in Sachs2
dGVGsQDSLITQk	scaffold655.g8	scn4aa	-3.5428753	-1.8139347	0.33433748	1.34649251	-0.1790383	-1.5879124	0.15925255	0.73755484	no	reduction in Sachs2
ePSSVkLsTEEQR	scaffold655.g9	scn4aa	NA	NA	1.03824938	1.62589816	-1.2962605	-2.2604421	0.67158301	0.84100052	no	missed cleavage. No muscle detected. Reduction in both Sachs' EOs, but >2 fold in Sachs2.
iVDGITNcVESPtLNLPIVk	scaffold655.g9	scn4aa	NA	NA	1.04916173	0.42067524	-0.4849104	-0.5963045	0.3622831	0.72441573	no (not localized)	reduced in main2.
kAsLASQLTqNQEAEITDDGDDAIk	scaffold655.g9	scn4aa	-1.9226716	-0.9221922	1.15180658	1.35663765	0.23640364	0.55806526	-0.2828816	-1.4399135	no	missed cleavage. reduced in hunters2
kPNTsVLFPkPNR	scaffold716.g5	chrnb1	0.00933777	-0.3350268	0.49868972	0.49119774	-0.6867629	-1.2011657	-0.0093986	0.27170127	no (not localized)	missed cleavage. nearly 2 fold decreased in sachs2.
sELMFEkQsER	scaffold118.g27	chrnd	-0.5160777	-0.8373741	1.07249399	0.95054455	-1.3664106	-1.4862205	0.37932109	0.52640809	no	missed cleavage. reduced in both sachs'
tGNPNINVDGSIDsR	scaffold259.g13	ache	-0.5027569	-0.6279032	0.78271699	0.62465888	-0.0248969	0.43603787	0.02447457	-1.0564471	no	slight decreases in main, slight increases in sachs2, large decrease in hunters2.
hVGLNTDsLk	scaffold259.g13	ache	-1.1973172	-0.7199081	1.23800252	1.65229187	-1.1314147	-1.580149	0.62623549	0.47805428	no	decrease in Sachs' EO (both)
PROTEIN LEVEL:												
	scaffold672.g4	atp1a2a	-1.2932605	-0.5760177	0.50402484	1.41393316	-0.0921845	-0.3245595	0.08664625	0.26478604		
	scaffold610.g4	scn4ab	-0.6160314	-0.6892974	0.77867838	1.61281318	-0.3645635	-0.5946522	0.29077372	0.41985803		
	scaffold655.g8	scn4aa	-0.9514262	-0.7807098	0.75511116	1.2281868	-0.317845	-0.2825173	0.26030592	0.23614946		
	scaffold655.g9	scn4aa	-1.6708021	-1.1233131	0.74955083	1.6102026	-0.5654256	-0.6473671	0.40516892	0.44525589		
	scaffold716.g5	chrnb1	-0.793777	-0.1524882	0.66122265	0.99514047	-0.2731819	-0.3425646	0.22959714	0.13790032		
	scaffold118.g27	chrnd	-0.0669639	0.0188425	0.39443979	0.7204699	-0.6039736	-0.2372741	0.06399312	-0.0190919		
	scaffold259.g13	ache	-1.4894829	-0.6386657	0.93881707	1.44882411	-0.4002145	-0.4116901	0.31296062	0.31991589		

Table 5.S3. Phosphopeptides in EOD-related proteins that differ in abundance compared to protein abundance. Shown in this table are the log2 (tissue/median) values of each differentially abundant phosphopeptide in all eight tissues, as well as the protein abundance values for all tissues (bottom of table). Abundance differences were considered significant if the difference between the phosphopeptide abundance and the protein abundance was at least two fold (difference in $\log_2$ values of at least 1). Orange highlighted rows indicate other phosphopeptides that differ in abundance, but have missed cleavages, and so were not discussed further.

Large Supplemental Files (available by shared Box link, online):

Supplemental Table 5.S4 (.XLS): Expression values (RNA) for all predicted genes in assembly. Expression values are in "reads per kilobase transcript", as described in the methods. Raw reads from [8], and include brain, spinal cord, whole heart, skeletal muscle, main EO, Sachs' EO, Hunter's EO, and whole kidney.

Supplemental Table 5.S5 (.XLS): Raw output from Proteome Discoverer, unenriched whole proteome samples. Values are unnormalized raw channel intensities, for input into custom script for normalization and quantitation.

Supplemental Table 5.S6 (.tsv): Median normalized channel intensity values for unenriched, whole proteome samples.

Supplemental Table 5.S7 (.XLSX): Intensity ratios, log2 (tissue/median), on a per protein group basis.

Supplemental Table 5.S8 (.XLSX): Raw output from Proteome Discoverer, titanium dioxide enriched phosphopeptides.

Supplemental Table 5.S9 (.TSV): Median-normalized channel intensity values for titanium dioxide enriched phosphopeptides.

Supplemental Table 5.S10 (.XLSX): Normalized intensity ratios, log2(tissue/median, on a per peptide basis. (A) All peptides identified (B) Only phosphopeptides.

Supplemental Table 5.S11 (.TSV): Novel and known phosphosites in *E. electricus* proteins. Contains phosphosite information for a subset of proteins discussed in the manuscript. A phosphosite was considered localized if it had a localization score of 75% or greater. Whether a phosphosite was considered known in mammals or novel was determined based on protein alignments with *E. electricus*, human, mouse, and zebrafish sequences. A phosphosite in *E. electricus* was considered novel if it was at least five amino acids away from a known phosphosite (this information is recorded in the "notes" column).

Supplemental Table 5.S12 (.XLSX): Correlation of RNA and protein abundance values. All abundance values for RNA or protein are in log2 (EO/muscle). "Group" values in each table have following meanings: 1: not DE in RNA or protein, 2: DE in protein only, 3: DE in RNA only, 4: DE in both RNA and protein, and 5: DE in both RNA and protein, but in opposite directions. (A) Eel1 correlation values, RNA expression values and protein abundance values, for all three EOs. (B) Eel 2 correlation values, RNA expression values, and protein abundance values, for all three EOs. (C) main EO, the join between tabs 1 and 2, where in both biological replicates, the gene models showed the same pattern (D) Join for Sachs' EO, (E) Join for Hunter's EO.

References

1. Bennett MVL: Electric Organs. In: *Fish Physiology.* Edited by Hoar WS, Randall DJ, vol. Volume 5: Academic Press; 1971: 347-491.
2. Moller P: Electric organs. In: *Electric Fishes: History and Behavior.* London: Chapman & Hall; 1995.
3. Albert JS, Crampton, W.G.R.: Electroreception and electrogenesis. Pp. 431-472 in The Physiology of Fishes, 3rd Edition. . In. Edited by Claiborne DHEaJB. Boca Raton, FL: CRC Press; 2005.
4. Coates CW, Cox RT: A comparison of length and vlotage in the electric eel, *Electrophorus electricus* (Linnaeus). *Zoologica; scientific contributions of the New York Zoological Society* 1945, 30:89-93.
5. Catania K: The shocking predatory strike of the electric eel. *Science* 2014, 346(6214):1231-1234.
6. Finger S, Piccolino M: The shocking history of electric fishes : from ancient epochs to the birth of modern neurophysiology. New York: Oxford University Press; 2011.
7. Altelaar AF, Munoz J, Heck AJ: Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature reviews Genetics* 2013, 14(1):35-48.
8. Gallant JR, Traeger LL, Volkening JD, Moffett H, Chen PH, Novina CD, Phillips GN, Jr., Anand R, Wells GB, Pinch M *et al*: Nonhuman genetics. Genomic basis for the convergent evolution of electric organs. *Science* 2014, 344(6191):1522-1525.
9. Mate SE, Brown KJ, Hoffman EP: Integrated genomics and proteomics of the Torpedo californica electric organ: concordance with the mammalian neuromuscular junction. *Skeletal muscle* 2011, 1(1):20.

10. Nazarian J, Hathout Y, Vertes A, Hoffman EP: The proteome survey of an electricity-generating organ (Torpedo californica electric organ). *Proteomics* 2007, 7(4):617-627.

11. Gotter AL, Kaetzel MA, Dedman JR: Electrophorus electricus as a model system for the study of membrane excitability. *Comparative biochemistry and physiology Part A, Molecular & integrative physiology* 1998, 119(1):225-241.

12. Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E: PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic acids research* 2015, 43(Database issue):D512-520.

13. Munjaal RP, Connor CG, Turner R, Dedman JR: Eel electric organ: hyperexpressing calmodulin system. *Molecular and cellular biology* 1986, 6(3):950-954.

14. Ben-Johny M, Yang PS, Niu J, Yang W, Joshi-Mukherjee R, Yue DT: Conservation of Ca2+/calmodulin regulation across Na and Ca2+ channels. *Cell* 2014, 157(7):1657-1670.

15. Wang C, Chung BC, Yan H, Lee SY, Pitt GS: Crystal structure of the ternary complex of a NaV C-terminal domain, a fibroblast growth factor homologous factor, and calmodulin. *Structure* 2012, 20(7):1167-1176.

16. Arnegard ME, Zwickl DJ, Lu Y, Zakon HH: Old gene duplication facilitates origin and diversification of an innovative communication system--twice. *Proceedings of the National Academy of Sciences of the United States of America* 2010, 107(51):22172-22177.

17. Markham MR: Electrocyte physiology: 50 years later. *The Journal of experimental biology* 2013, 216(Pt 13):2451-2458.

18. Ching B, Woo JM, Hiong KC, Boo MV, Choo CY, Wong WP, Chew SF, Ip YK: Na+/K+-ATPase alpha-subunit (nkaalpha) Isoforms and Their mRNA Expression Levels, Overall Nkaalpha Protein Abundance, and Kinetic Properties of Nka in the Skeletal Muscle and Three Electric Organs of the Electric Eel, Electrophorus electricus. *PloS one* 2015, 10(3):e0118352.

19. Lowe J, Araujo GM, Pedrenho AR, Nunes-Tavares N, Ribeiro MG, Hasson-Voloch A: Polarized distribution of Na+, K(+)-ATPase alpha-subunit isoforms in electrocyte membranes. *Biochimica et biophysica acta* 2004, 1661(1):40-46.

20. Shenkel S, Sigworth FJ: Patch recordings from the electrocytes of Electrophorus electricus. Na currents and PNa/PK variability. *The Journal of general physiology* 1991, 97(5):1013-1041.

21. Keynes RD, Martins-Ferreira H: Membrane potentials in the electroplates of the electric eel. *The Journal of physiology* 1953, 119(2-3):315-351.

22. Thornhill WB, Watanabe I, Sutachan JJ, Wu MB, Wu X, Zhu J, Recio-Pinto E: Molecular cloning and expression of a Kv1.1-like potassium channel from the electric organ of Electrophorus electricus. *The Journal of membrane biology* 2003, 196(1):1-8.

23. Markham MR, McAnelly ML, Stoddard PK, Zakon HH: Circadian and social cues regulate ion channel trafficking. *PLoS biology* 2009, 7(9):e1000203.

24. Mihaylova MM, Shaw RJ: The AMPK signalling pathway coordinates cell growth, autophagy and metabolism. *Nat Cell Biol* 2011, 13(9):1016-1023.

25. Chida T, Ando M, Matsuki T, Masu Y, Nagaura Y, Takano-Yamamoto T, Tamura S, Kobayashi T: N-Myristoylation is essential for protein phosphatases PPM1A and

PPM1B to dephosphorylate their physiological substrates in cells. *The Biochemical journal* 2013, 449(3):741-749.

26. Sparmann A, van Lohuizen M: Polycomb silencers control cell fate, development and cancer. *Nature reviews Cancer* 2006, 6(11):846-856.

27. Shi Y, Li Y, Zhang D, Zhang H, Li Y, Lu F, Liu X, He F, Gong B, Cai L *et al*: Exome sequencing identifies ZNF644 mutations in high myopia. *PLoS genetics* 2011, 7(6):e1002084.

28. Werth M, Walentin K, Aue A, Schonheit J, Wuebken A, Pode-Shakked N, Vilianovitch L, Erdmann B, Dekel B, Bader M *et al*: The transcription factor grainyhead-like 2 regulates the molecular composition of the epithelial apical junctional complex. *Development* 2010, 137(22):3835-3845.

29. Wang Q, Margolis B: Apical junctional complexes and cell polarity. *Kidney international* 2007, 72(12):1448-1458.

30. Melotte V, Qu X, Ongenaert M, van Criekinge W, de Bruine AP, Baldwin HS, van Engeland M: The N-myc downstream regulated gene (NDRG) family: diverse functions, multiple applications. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 2010, 24(11):4153-4166.

31. Newton-Cheh C, Eijgelsheim M, Rice KM, de Bakker PI, Yin X, Estrada K, Bis JC, Marciante K, Rivadeneira F, Noseworthy PA *et al*: Common variants at ten loci influence QT interval duration in the QTGEN Study. *Nature genetics* 2009, 41(4):399-406.

32. Pfeufer A, Sanna S, Arking DE, Muller M, Gateva V, Fuchsberger C, Ehret GB, Orru M, Pattaro C, Kottgen A *et al*: Common variants at ten loci modulate the QT interval duration in the QTSCD Study. *Nature genetics* 2009, 41(4):407-414.

33. Quigley IK, Stubbs JL, Kintner C: Specification of ion transport cells in the Xenopus larval skin. *Development* 2011, 138(4):705-714.

34. Penttila S, Palmio J, Suominen T, Raheem O, Evila A, Muelas Gomez N, Tasca G, Waddell LB, Clarke NF, Barboi A *et al*: Eight new mutations and the expanding phenotype variability in muscular dystrophy caused by ANO5. *Neurology* 2012, 78(12):897-903.

35. Wahbi K, Behin A, Becane HM, Leturcq F, Cossee M, Laforet P, Stojkovic T, Carlier P, Toussaint M, Gaxotte V *et al*: Dilated cardiomyopathy in patients with mutations in anoctamin 5. *International journal of cardiology* 2013, 168(1):76-79.

36. Hnia K, Ramspacher C, Vermot J, Laporte J: Desmin in muscle and associated diseases: beyond the structural function. *Cell and tissue research* 2015, 360(3):591-608.

37. Costa ML, Mermelstein CS, Froes MM, Chagas C, Moura Neto V: Differences in the isodesmin pattern between the electric organs of Electrophorus electricus L. *Comparative biochemistry and physiology Part B, Biochemistry & molecular biology* 1998, 119(4):715-719.

38. Cordeiro MCR, Neto VM, Benchimol M, Faria MVC, Chagas C: Microheterogeneity of Desmin in the Electric Organ and Dorsal Muscle of the Electric-Eel Electrophorus-Electricus. *Comp Biochem Phys A* 1995, 111(3):345-350.

39. Paulin D, Li Z: Desmin: a major intermediate filament protein essential for the structural integrity and function of muscle. *Experimental cell research* 2004, 301(1):1-7.

40. Capetanaki Y, Papathanasiou S, Diokmetzidou A, Vatsellas G, Tsikitis M: Desmin related disease: a matter of cell survival failure. *Current opinion in cell biology* 2015, 32:113-120.

41. Esquibel MA, Miguens FC, Machado RD: Scanning electron microscopy of the electric organs of Electrophorus electricus L. II. Organs of Sachs and Hunter. *Cell and tissue research* 1985, 241(3):585-592.

42. Traeger LL, Volkening JD, Moffett H, Gallant JR, Chen PH, Novina CD, Phillips GN, Jr., Anand R, Wells GB, Pinch M *et al*: Unique patterns of transcript and miRNA expression in the South American strong voltage electric eel (Electrophorus electricus). *BMC genomics* 2015, 16:243.

43. Geering K: FXYD proteins: new regulators of Na-K-ATPase. *American journal of physiology Renal physiology* 2006, 290(2):F241-250.

44. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W: Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 2011, 27(4):578-579.

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

46. 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.

47. Slater GS, Birney E: Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics* 2005, 6:31.

48. Keller O, Kollmar M, Stanke M, Waack S: A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* 2011, 27(6):757-763.

49. Gurevich A, Saveliev V, Vyahhi N, Tesler G: QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013, 29(8):1072-1075.

50. Anders S, Pyl PT, Huber W: HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 2015, 31(2):166-169.

51. Anders S, Huber W: Differential expression analysis for sequence count data. *Genome biology* 2010, 11(10):R106.

52. Sugiyama N, Masuda T, Shinoda K, Nakamura A, Tomita M, Ishihama Y: Phosphopeptide enrichment by aliphatic hydroxy acid-modified metal oxide chromatography for nano-LC-MS/MS in proteomics applications. *Molecular & cellular proteomics : MCP* 2007, 6(6):1103-1109.

53. Hebert AS, Richards AL, Bailey DJ, Ulbrich A, Coughlin EE, Westphall MS, Coon JJ: The one hour yeast proteome. *Molecular & cellular proteomics : MCP* 2014, 13(1):339-347.

54. Gregory R. Warnes BB, Lodewijk Bonebakker, Robert Gentleman, Wolfgang Huber Andy Liaw,, Thomas Lumley MM, Arni Magnusson, Steffen Moeller, Marc Schwartz and Bill Venables: gplots: Various R Programming Tools for Plotting Data. In., 2.17.0 edn; 2015.

55. Kolde R: pheatmap: Pretty Heatmaps. In., 1.0.2 edn; 2015.

Chapter 6: Final Perspectives and Future Directions

"As story tells us, in the Indian Seas,

Of finny Race, the cold Torpedo plays;

Who to the Hook, by greedy Hunger brought,

The Fish and Fisher both at once are caught:

The Icy Venom from his Rod assails,

And soon thro' all his curdling Blood prevails…"

- *Universal Spectator and Weekly Journal, June 12, 1731*

Future Directions:

The experiments presented in this dissertation jolt the field of electric fish biology into the 'omics' era. However, there are many elements that remain unknown in electric fishes, especially *E. electricus.* The following three analyses, discussed below, are areas of research that I would find particularly interesting, and build off of the framework established in this dissertation.

Improving genome assembly of 3'UTR regions to perform miRNA target prediction

In Chapter 3 of this dissertation, we analyzed miRNA expression in eight tissues in *E. electricus*, including the three EOs and muscle, and identified several novel miRNAs and three miRNAs that are known to inhibit muscle development and are highly abundant in our three EOs. However, in this study, we were limited by the fragmentation of our genome assembly and the resulting gene models in it, in particular, regions of the genes encoding the 3' untranslated region (3'UTR) of transcripts. This is problematic for miRNA target prediction because miRNAs predominately regulate their targets by interacting with the

3'UTR region of transcripts, although miRNA binding to the 5'UTR and exonic regions has been described as an alternative and robust regulatory mechanism. Predicting the targets of the highly abundant and previously described miRNAs in EO would help inform us of their possible function in electrocytes.

In Chapter 5, I performed additional scaffolding and gap filling of the genome, generally improving the gene models predicted in it. Although I have not examined the quality of 3' UTRs in the new assembly, it may be improved enough to do such an analysis. Assuming the genome presented in Chapter 5 is of sufficient quality, all the pieces would be in place for such an analysis, and such programs exist to perform miRNA target enrichment [1]. If the genome presented in Chapter 5 is still too fragmented or incomplete for a robust miRNA target experiment, additional Pacific Biosciences sequences were generated for *E. electricus* that have not been included in any *E. electricus* genome assembly to date. Alternately, instead of focusing on the genome, one could spend time improving the transcriptome assembly of *E. electricus* or *Sternopygous macrurus* (a Gymnotiform for which we have a transcriptome assembly and have performed miRNA sequencing on muscle and EO), which are both currently difficult to work with given their massive nature (hundreds of thousands of individual transcripts in each assembly) but could be improved on (Chapter 3). Analyzing the possible targets of the novel miRNAs identified in *E. electricus* would be a potentially interesting follow up analysis, and ultimately lead to further hypotheses as to the development of electric organs from myogenic precursors.

<u>Identifying phosphorylation sites in *E. electricus* important for electrocyte function.</u>

Although in Chapter 5, we identified several novel phosphorylation sites across proteins important for electric organ discharge, including the acetylcholine receptor, the voltage-gated sodium channel, and the Na+/K+-ATPase, a significant goal is to identify which phosphorylation sites are serving a biological function. Because we are unable to use genetic approaches at understanding the role of each phosphorylation site in *E. electricus,* I began experiments aimed at teasing apart which phosphorylation sites may be more important than others using an approach that treated slices of main EO with various compounds *ex vivo* (see Appendix I for experimental details). For five minutes each, I treated tissue slices with an agent to inhibit the voltage-gated sodium channel (tetrodotoxin, the pufferfish toxin), an agent to inhibit the Na+/K+-ATPase (ouabain), or an agent to activate the acetylcholine receptor (carbamylcholine). The purpose of this experiment will be to compare changes in protein phosphorylation, in particular to the three aforementioned proteins/protein complexes, in each treatment. If treatment results in phosphorylation changes at distinct phosphorylated residues, it may be evidence for the biological importance of a given phosphorylation site.

An alternative approach to this would be to consider protein phosphorylation in an additional gymnotiform species (such as *S. macrurus* or *E. virescens*). We have assembled transcriptomes from both of these species (Chapter 4), although the assembly for each could be improved prior to using them as a protein database for mass spectral searching, due to the fact the assemblies are massive (hundreds of thousands to millions of independent transcripts). Novel phosphorylation sites that were shared among different Gymnotiformes tested could give some weight to certain phosphorylation sites having

biological importance over others. Further, conserved phosphorylated residues that are mutated to

Gymnotiformes as models for tissue regeneration

Many vertebrate fish species hold a remarkable ability to regenerate tissue post amputation or injury; among these, Gymnotiformes (notably weakly electric Gymnotiformes), show a remarkably high capacity for tissue regeneration. Post tail amputation, Gymnotiformes are able to completely regenerate all tissues lost (spinal cord, skeleton, skeletal muscle, and electric organs, to name a few), and can do so without the formation of a scar (reviewed in [2]), and some can do inexhaustibly following repeated tail amputations. This is in contrast to zebrafish, which is able to regenerate many tissues and has many benefits because it is a model system, but is unable to regenerate tail post amputation. Among all Gymnotiformes, the species that has been studied the most with respect to tissue regeneration post tail amputation is *S. macrurus*. In this system, it has been demonstrated that skeletal muscle and EO are replaced post tail amputation from a population of myogenic progenitors (satellite cells) found in the adjacent, intact tissue [3], although it is unclear whether muscle and EO have distinct progenitor cells that give rise to each tissue, respectively, or whether it's the same population of progenitors that give rise to both muscle and EO [2]. All studies of regeneration in Gymnotiformes have done so without the presence of a genome or transcriptome sequence assembly. Improving and using the transcriptome assembly of *S. macrurus* would enable studies at the transcript or protein level across time points during the development of the blastema (mass of

undifferentiated cells that appears at end of wound site post tail amputation), and throughout tail regeneration.

Along similar lines, the electrocyte phenotype is dependent on neural input, and when this neural connection is severed in *S. macrurus*, electrocytes phenotypically appear more muscle-like (for example, denervated electrocytes begin to express sarcomeric proteins over the course of weeks) [4]. The underlying mechanism of this, or an understanding of how innervation influences electrocyte-specific gene expression is unknown. Using the transcriptome assembly of *S. macrurus*, one could explore changes in transcript or protein abundance in electrocytes post denervation, to get an understanding of how neural input is controlling or influencing the electric organ phenotype.

Final Thoughts

Electric fishes have been historically important across several scientific disciplines. Despite their importance, prior to the onset of this thesis dissertation, no study has previously characterized electric organs at a large-scale molecular level. The data presented throughout this dissertation will be useful to many fields, and enable the ability to ask specific biological questions that were previously inaccessible due to the lack of a genome or transcriptome for any electric fish species.

Finally, strong-voltage electric fishes, including *E. electricus,* the electric rays (*Torpedo*), and the electric catfish have shaped (and continue to shape) our culture, inspiring stories, poetry, and paintings (Figure 6.1). It is my personal experience that the topic of the electric eel has an amazing ability to bring out a child-like curiosity in everyone.

I will always look back on my years as “the electric eel woman” with a smile, and am grateful to have had this wonderful experience.

Figure 6.1. Ode to Ollie. Painted by Cheryl Redman (UW Biotechnology Center).

References:

1. Maragkakis M, Alexiou P, Papadopoulos GL, Reczko M, Dalamagas T, Giannopoulos G, Goumas G, Koukis E, Kourtis K, Simossis VA *et al*: Accurate microRNA target prediction correlates with protein repression levels. *BMC bioinformatics* 2009, 10:295.
2. Unguez GA: Electric fish: new insights into conserved processes of adult tissue regeneration. *The Journal of experimental biology* 2013, 216(Pt 13):2478-2486.
3. Weber CM, Martindale MQ, Tapscott SJ, Unguez GA: Activation of Pax7-positive cells in a non-contractile tissue contributes to regeneration of myogenic tissues in the electric fish S. macrurus. *PloS one* 2012, 7(5):e36819.
4. Unguez GA, Zakon HH: Reexpression of myogenic proteins in mature electric organ after removal of neural input. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 1998, 18(23):9924-9935.

Appendix I*: Ex vivo* treatment of main electric organ slices

Purpose:

The objective of this experiment was to determine changes in protein phosphorylation after five minute exogenous *ex vivo* treatment of main EO slices with various compounds, to determine whether any sites change post inhibition/activation. Any changing phosphorylation sites may be indicative of functional phosphorylation in our proteins of interest, namely, the acetylcholine receptor, the voltage-gated Na+ channels, and the Na+/K+-ATPase

Materials and Methods:

A total of 500 mL of Ringer solution was prepared as previously described for *E. electricus* [1] (169 mM NaCl, 5 mM KCl, 3 mM $CaCl_2$, 1.5 mM $MgCl_2$, 1.2 mM Na_2HPO_4, 0.3 mM NaH_2PO_4). Ringer solution containing treatment compounds were prepared immediately before beginning experiment, stored in 50 mL conical tubes until use (see below Table Ap1). All animals were euthanized one at a time and followed our animal protocol (ice bath for 10 minutes followed by swift decapitation). Slices of main EO ~0.4-0.5 cm in thickness along length of tail corresponding to main EO section were created and kept wrapped loosely in plastic wrap and on ice until ready for use. Two EO slices at a time were received five minute treatment: They were removed from the plastic wrap and immersed completely in a square petri plate containing Ringer solution and treatment where they stayed for five minutes. After five minutes of treatment, tissue was removed from petri plate, the dorsal-end of the sample was removed (containing muscle, swim bladder, vertebrae, etc) and flash frozen in liquid nitrogen. Slices transferred to 50 mL conical tube and stored at -80 until use. The order of treatment varied between the two

eels used in the study, but two controls flanked the three treated samples—the first two tissue slices were immersed in Ringer solution only for five minutes followed by flash freezing (control 1), and the last two slices were treated the same way (control 2). This was done to control for any changes that may be happening in the EO tissue as it sat on ice over time. This was repeated for two eels in one day.

Protein extraction:

In total ten samples were processed- two fish, and five samples per fish (two controls and three treatments). Pieces of main EO were removed from EO slice (~ 0.3-0.4 g tissue) and pulverize on liquid nitrogen using a mortar and pestle. Three mL of 8M urea lysis buffer containing protease and phosphatase inhibitors was added, as previously described (Chapter 5 methods). Samples were homogenized using a bead beater, splitting each sample across three Eppendorf tubes containing a large steel bead and homogenized at 20Hz for three minutes at 4°C. Post homogenization, samples were transferred to a 50 mL conical tube for methanol chloroform extraction, followed by protein resolubolization, and quantitation (as previously described, see Chapter 5 methods). A total of 500 ug protein was incubated with dithothreitol (DTT) at a final concentration of 5 mM for 35 minutes at 65°C. Alkylation was performed by incubating with idoacetamide at a final concentration of 12.5 mM at room temperature in the dark. Reactions were quenched by adding DTT to a final concentration of 10 mM. Samples were digested with trypsin and lys-c at a final ratio of 1:100 enzyme to protein for each enzyme, at 37°C overnight.

Digests were acidified by adding formic acid to a final concentration of 0.5 mM. Samples were cleaned up using solid phase (C18) chromatography (1cc column, SepPack,

Waters), following protocol in Chapter 5 methods. The eluate volumes were measured and divided such that 400 ug of sample would be used in final labeling reaction, and 100 ug would be separated from it for quality checks before labeling reaction. Samples were dried and stored at -80°C until use.

Labeling reaction:

The ten 400 ug samples were each labeled with one tube of 0.8 mg TMT-10 plex tag, the following tags assigned to each sample: from the first fish, control-1 (TMT-126), ouabain (TMT 127N), carbamylcholine (TMT 127C), tetrodotoxin (TMT 128N), control-2 (TMT 128C), and from the second fish control-1 (TMT 129N), ouabain (TMT 129C), carbamylcholine (TMT 130N), tetrodotoxin (TMT 130C), control-2 (TMT 131). Labeling reactions followed manufacturers protocol with the following exceptions: The lyophilized peptide samples were brought up in 100 ul of 200 mM TEAB. (instead of 100 mM TEAB). Labeling reactions were performed for 2.5 hours at room temperature on a shaking table, and vortexed every ~30 minutes. For testing ratios, 5 ul of each sample was pulled from each labeling reaction and pooled in 1:1:1:1:1:1:1:1:1:1 ratio based on initial protein concentration with the other samples. The remaining sample was froze at -80°C.

Samples were processed as described in chapter 5. Briefly, the TMT test mix was desalted on a C18 column (Omix tip, brand), and desalted samples were run on the Orbitrap Elite for ratio testing. Peptides were combined in a in 1:1:1:1:1:1:1:1:1:1 ratio based on the summed reporter ion intensity for each channel. The sample was prefractionated using a high pH reversed-phase method. The resulting 30 second fractions were pooled into 1 minute fractions, and then every fourth fraction was combined. These resulting fractions, four in total, were lyophilized, and then desalted on a C18 column

(SepPack, Waters, 3cc). Phosphopeptides were enriched from these four fractions, using the glycolic acid method described in Chapter 5. These samples are ready for data acquisition on the Orbitrap Fusion, for analysis of phosphosites that change in the main EO tissue post five minutes of treatment with inhibitory or excitatory compounds.

Compound	**Concentration**	**Ref. for conc**	**Vendor**
Tetrodotoxin	7x10-8 M	[2]	Tetrodotoxin 1mg. Fischer Scientific (NC0066215)
Ouabain	0.01 M	[3]	Ouabain Octahydrate 98% 1G from VWR (AAJ60724-03)
Carbamylcholine chloride	10^{-4} M	[4]	Carbamylcholine chloride, 5g. VWR (200055-290)

Table Ap1. Compound concentrations used in this experiment.

References:

1. Keynes RD, Martins-Ferreira H: Membrane potentials in the electroplates of the electric eel. *The Journal of physiology* 1953, 119(2-3):315-351.
2. Dettbarn WD, Higman H, Rosenberg P, Nachmansohn D: Rapid and reversible block of electrical activity by powerful marine biotoxins. *Science* 1960, 132(3422):300-301.
3. Lowe J, Araujo GM, Pedrenho AR, Nunes-Tavares N, Ribeiro MG, Hasson-Voloch A: Polarized distribution of Na+, K(+)-ATPase alpha-subunit isoforms in electrocyte membranes. *Biochimica et biophysica acta* 2004, 1661(1):40-46.

4. Lester HA, Changeux JP, Sheridan RE: Conductance increases produced by bath application of cholinergic agonists to Electrophorus electroplaques. *The Journal of general physiology* 1975, 65(6):797-816.