January 2019

Aging Neuro-Behavior Ontology

Fernando Martínez-Santiago [a,1], M. Rosario García-Viedma [e], John A. Williams [b,c,d], Luke T. Slater [b,c] and Georgios V Gkoutos [b,c]

[a] *SINAI Group, Computer Science Department, University of Jaén, SPAIN*
[b] *Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK*
[c] *Institute of Translational Medicine, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK*
[d] *MRC Harwell Institute, Mammalian Genetics Unit, Harwell Campus, UK*
[e] *Psychology Department, University of Jaén, SPAIN*

Abstract. It is known that the aging process entails a cognitive decline in certain processes such as attention, episodic memory, working memory, processing speed and executive functions. In recent years, efforts have been made to investigate the potential of Information and Communication Technologies to improve cognitive functioning and quality of life in older adults with and without cognitive impairments. In this paper, we propose the Aging Neuro-Behaviour Ontology (ANBO), a formal model of cognitive processes involved in day-to-day living and whose performance usually decline with age. ANBO has been created with the aim of being an aid in developing tools for cognitive rehabilitation by means of integrating aging-related behaviors with monitoring activities of daily living. As an example of these tools, we introduce an integration of ANBO with the Ontology SmartLab Elderly (OSLE), an ontology related with Telehealth Smart Homes wherein activities of daily living are recorded. This ANBO and OSLE integration enable the interpretation of these activities as the result of cognitive processes of interest in the domain of elderly decline.

Keywords. elderly, neuro-psychology,, ontology engineering, ambient intelligence

1. Introduction

Cognition is a collection of brain-based functions which allow us to interact successfully with the world around us. These cognitive abilities are necessary to carry out any activity of daily living (ADL), from the simplest to the most complex tasks. Thus, daily tasks can be defined by the degree to which they require specific cognitive operations and it is possible classify functional tasks based on required cognitive abilities (Lawton and Brody, 1970; Farias et al., 2008). For instance, driving is a task that takes place in complex environments and it requires organizing and implementing various sub-tasks such as controlling speed, interpreting abstract signs, planning the best route, and planning alternative routes in the case of unforeseen events. Even the simplest activities, such as answering the telephone, require the programming and use of different cognitive operations: per-

[1] Corresponding Author: Fernando Martínez-Santiago; E-mail: dofer@ujaen.es.

ception (hearing the ring tone), decision making (answering or not), motor skills (lifting the receiver), language (producing and understanding language), and social skills (interpreting tone of voice and interacting properly with another human being). Some of the main cognitive capacities are:

- **Perception** permits recognition and interpretation of sensory stimuli (smell, touch, hearing,etc) (Bruce and Young, 1986; Humphreys and Riddoch, 2017; Young and Bruce, 2011).
- **Attention** is a complex mechanism with several different facets (focused, sustained, selective, alternating and divided). This capacity is considered a central control mechanism and its principal function is to direct and guide the conscious activity of the organism according to a specific goal or objective (Posner and Dehaene, 1994).
- **Memory**. Learning and human memory are part of a complex processing system (short-term/working memory - limited storage, long-term memory - unlimited storage, implicit and explicit memory) that is in charge of coding, storing, building, reconstructing and recovering perceptions, knowledge, facts, abilities, emotions, plans, etc (Atkinson and Shiffrin, 1968; Baddeley, 1992, 2000; Tulving and Craik, 2000).
- **Language** forms one of the very complex human capacities that allow us to communicate thoughts, ideas, feelings, doubts, desires and needs (García-Viedma and Fernández-Guinea, 2010; Patterson and Shewell, 2013).
- **Executive functions** are abilities that allow us to transform our thoughts into actions efficiently. According to Lezak (1995) these can be grouped into the following components: formulations of goals, planning processes and strategies to achieve objectives (to establish steps or action items, to evaluate different alternatives), monitoring and supervision of action (recognition of achievement or non-achievement, flexibility or capacity for quickly switching to the appropriate mental mode), and efficient performance of the plans (control, correction, self-regulation of time and other qualities of execution).

Each cognitive ability plays an important part in the processing of new information. If one of these processes has deficits, the correct functioning of the other operations will be affected, and these impairments lead to difficulties in carrying out ADLs. Loss of cognitive functions causes interference with social ad occupational functioning. Consequently, functional independence and quality of life may be diminished.

The normal aging process entails changes in cognition. Although some cognitive functions, including world knowledge, verbal abilities and implicit memory are preserved in older adults (Ballesteros et al., 2013; Park et al., 2002; Wiggs et al., 2006), this aging process is associated with an important decrease in certain cognitive functions such as attention, episodic memory, working memory, processing speed, and executive functions (Rönnlund et al., 2005; Salthouse, 2010). These deficits do not have a significant impact on ADLs and functional independence, but these difficulties often progress into more serious conditions such as Mild Cognitive Impairment (MCI) or dementia, principally Alzheimer's disease (AD; Brown et al., 2011; Jekel et al., 2015). Detecting functional decline along the continuum from normal aging to dementia is crucial because these difficulties are considered a risk factor and prodromal stage of the dementia. Assessment of daily living activities would also allow to identify abnormal behaviors as

indicators of cognitive decline and it is an important aspect of neuropsychological evaluation. So that some functional instruments have been development to measure deficit and change in domains of everyday functioning relevant to specific cognitive domain, such as ECog (Farias et al., 2008).

The cognitive changes are evident in aging process. However, previous findings show that cognitive training interventions and programs can improve cognition in healthy older adults (Ball et al., 2002). Thus, training cognitive abilities delays cognitive decline and helps older individuals to maintain independence and a higher quality of life for longer.

ANBO is an effort to facilitate the systematic representation of behavior and behavioral phenotypes related to the elderly, with special interest in enabling computational access and reasoning by means of these representations in knowledge-based systems. As a first case of use, we propose the application of ANBO to a specific domain: Teleheath Smart Homes (Latfi et al., 2007). In the context of this work, the role of ANBO is to provide models to monitor cognitive skills and their phenotypes, whose performance usually decreases with age.

The rest of this paper is organized as follows. Section 2 presents the cognitive aging process, its main characteristics and the limitations of traditional methods in assessment and intervention, as well as recent technological advances in relevant methodologies. Section 3 describes ANBO, our proposal to model particularly sensitive cognitive processes over a period of years. Section 4 introduces some of the ontologies related to ANBO because they are reused as part of ANBO as well as sharing a similar topic of interest. Section 5 depicts the way that OSLE and ANBO are integrated while maintaining the independence of both entities. Section 6 evaluates the design and implementation of ANBO. The paper finishes with conclusions and proposals for future work.

2. Assessment and Training of Cognitive Abilities

The global population aged 60 years or over numbered 962 million in 2017 and it is estimated to grow to about 2 billion by the year 2050. In Europe, the population aged 60 years or over reached 183 million in 2017 and the number of older persons is projected to grow to 247 million in 2050 (United Nations and Social Affairs, 2017). Aging is a complex process in which there are cognitive changes affecting a large population. This process is associated with a decline in certain cognitive functions such as attention, episodic memory, working memory, processing speed, and executive functions (Rönnlund et al., 2005; Salthouse, 2010). Generally, cognitive decline has no significant impact on functional independence for older adults, but these difficulties often progress into more serious conditions like dementia (a neurodegenerative illness that is characterized by a decline in mental ability severe enough to interfere with tasks on a day-to-day basis) and Alzheimer's disease (AD). Today, 50 million people are living with dementia and this number will increase to 152 million in 2050, a 204 percent increase which is highly significant (World Health Organization, 2017). Given the overall aging of the population and the cognitive decline associated therewith, there is a great deal of interest in the maintenance of cognitive functions and independent living in older adults for as long as possible. Several studies show that cognitive training interventions and programs can improve cognition in healthy older adults (Ball et al., 2002; Karbach and Kray, 2009).

These studies indicate that neuroplasticity (the physiological capacity of the brain to form and strengthen neuronal connections) is not limited to the early years but extends into old age (Pascual-Leone et al., 2005; Willis et al., 2006). Moreover, cognitive intervention could be beneficial to the treatment of typical and atypical cognitive aging (Engvig et al., 2010) and may reduce the risk of dementia (Verghese et al., 2003). On the other hand, an effective cognitive intervention requires a careful evaluation of cognitive and functional status. That is to say, it is necessary to assess different cognitive operations and how these support everyday functioning. The assessment is typically based on validated hand-written neuropsychological tests or scales, which have shown to be excellent tools for evaluating specific cognitive processes. However, these instruments were not designed to evaluate functional status and the results do not always reflect problems experienced in everyday life (Pugnetti et al., 1998; Mcalister et al., 2016). Thus, these tests present a poor ecological validity, as results from neuropsychological test are not easily generalizable to real-world functioning (Mcalister et al., 2016; Valladares-Rodríguez et al., 2016). In a similar vein, the ecological validity of the actual rehabilitation activities has been questioned, as well as the generalization of new abilities, knowledge and/or skills (Rizzo et al., 2004). More ecologically valid assessment and rehabilitation scenarios are needed. Realistic environments provide high ecological validity because these imply tasks that everyone can find in everyday life. Information and Communication Technologies (ICTs) are tools which could provide scenarios that mimic the real world or are integrated into real-world environments. This would enable evaluation and rehabilitation without the limitations of traditional methods.

3. Description of ANBO

The main purpose of ANBO is not to provide a psychological model of cognition or perception, but a formal representation of such processes. With this in mind, the focus is not on how a cognitive process works, but (i) the conditions under which a process could be triggered, (ii) which results are expected to be achieved at the end of the process, and (iii) which qualities define typical or atypical manifestations of the process. For example, regarding the *visual search*[2] process, ANBO does not say anything about how the process actually happens but rather gives a specification of prerequisites and expected outcomes of this process such as:

- The visual search process is focused on *physical qualities* of objects such *shape*, *size* and *color*.
- A *visual stimulus* is necessary, related to one or more *physical qualities*
- A *visual system* for perceiving the stimulus is necessary.
- Needing too much time to accomplish the task or picking up a wrong object are clues of an abnormal visual search.

[2] visual search, or alternatively visual search, is a type of perceptual task requiring attention that typically involves an active scan of the visual environment for a particular object or feature (the target) among other objects or features (the distractors). Visual Search is defined as an specialization of Visual Behaviour, that is defined as "Behavior related to the actions or reactions of an organism in response to a visual stimulus" in NBO and GO ontologies. Consequently, visual search is a kind of reaction (searching the desired object) as response to a specific visual stimulus (the objects or distractors to be "scanned").

ANBO is designed to support ontology reusability. We embraced the OBO Foundry (Smith et al., 2007) as a design framework. OBO proposes a set of principles including open use, collaborative development, non-overlapping and strictly-scoped content, and common syntax and relations. NBO itself is an OBO Foundry ontology, and our extensions follow these guidelines too.

3.1. Representative elements of ANBO

ANBO processes are made up of a number of constraints, behavior qualities and expected results. Processes explicitly modeled are given below in Table 1:

Cognitive function	ANBO Term ID	Related with
eye-hand coordination	NBO:0000341	perception
visual search	ANBO:0000004	perception
spatial orientation	ANBO:0000001	attention
focused/sustained/alternating/divided attention	ANBO:0000102-3/NBO:0000457-8,60	attention
working memory	NBO:0000180	memory behavior
semantic memory	NBO:0000186	long-term memory
problem solving	NBO:0000297	executive function
monitoring	ANBO:0000101	executive function
planning	ANBO:0000100	executive function
motor coordination	NBO:0000339	praxias

Table 1. Cognitive functions that are defined as behavioral processes in ANBO

Constraints constitute the input of a given cognitive process. These must be satisfied in order to make it possible to trigger the corresponding process. We distinguish between two different types of constraints (see Figure 2):

- Perception. The user has to have the capacity to perceive relevant information from their environment by means of one or more sensory systems. For example, the behaviour *visual search* is triggered in response to *visual perception* by means of a *visual system*.
- Focus. Each process has an environmental event or object of interest which is the focus of that process. For example, a pair of socks that are located in a drawer of a bedside table is the focus of putting clothes away.

A behavioral phenotype is an observed manifestation of a behavioral process. From the point of view of ANBO, a phenotype is one or more qualities of interest with (a) certain value/s as a result of the execution of the given cognitive process under particular conditions(Gkoutos et al., 2005). For example, if the process *visual search* consumes more time than usual (this has an increased duration) or a person becomes disoriented as a consequence of a malfunction of the spatial orientation process.

4. Related ontologies

ANBO expands or reuses a number of ontologies, inter alia, NBO (Gkoutos et al., 2012), PATO (Gkoutos et al., 2005), Uberon (Mungall et al., 2012) and GO (Consortium, 2004).

In this section we provide a brief description of these ontologies and the way that these ontologies are integrated into ANBO. Several of these ontologies are very large. As they are made up of several thousands of classes and relations, a partial export of relevant terms from each ontology is performed using OntoFox[3]. Ontofox is a web-based tool which facilitates ontology reuse, by allowing users to extract properties, annotations, and classes from ontologies and export. them. Ontofox follows and expands the Minimum Information to Reference an External Ontology Term (MIREOT) principle (Courtot et al., 2011).

4.0.1. Neuro-Behaviour Ontology (NBO)

NBO (Gkoutos et al., 2012) is the base upon which ANBO was built. NBO provides formal definitions and systematic representations of the processes involved in behavioral mechanisms and their related behavioral manifestations. ANBO is a specialization of NBO, emphasizing a formal representation of cognitive skills and their phenotypes. Thus, NBO provides the following elements to ANBO:

1. A taxonomy of cognitive processes. Eight of the eleven cognitive processes modelled in ANBO were directly imported from NBO. The other three concepts are defined within the new ontology as specializations of more general NBO concepts (see Figure 1).
2. The formal description of semantic relationships between concepts. By means of these relations is possible to make certain inferences - particularly, which

[3]http://ontofox.hegroup.org

Figure 1. ANBO taxonomy. Shaded squares filled with vertical lines represent processes native to NBO, and those with shaded horizontal lines represent processes which were added specifically to ANBO. Each grey shaded process in directly modeled with ADLs by the combination of ANBO and OSLE.

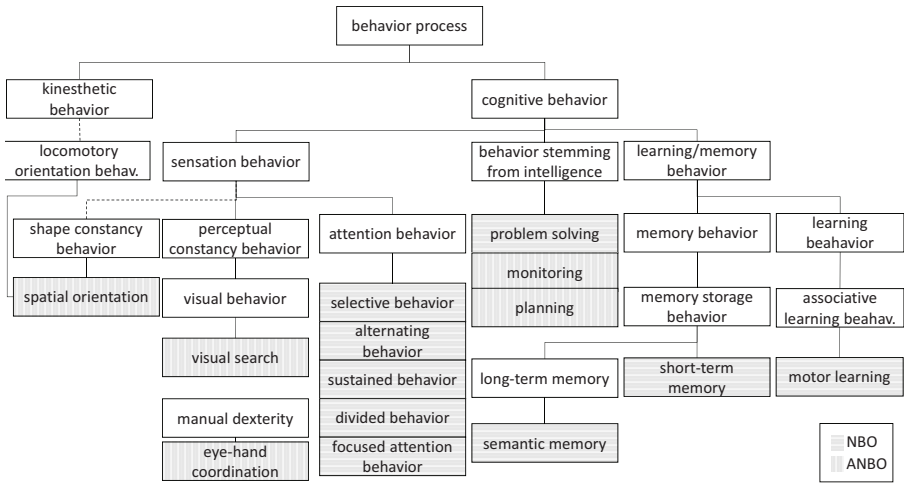
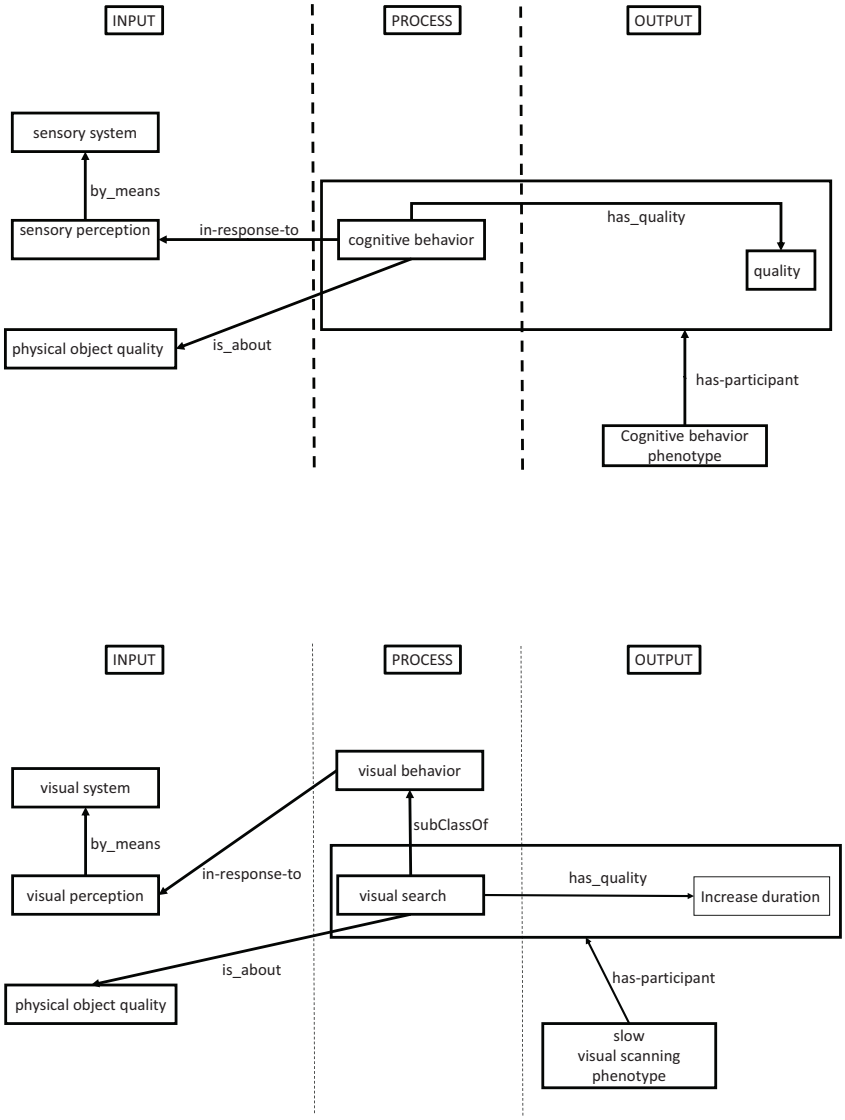


Figure 2. General structure of a cognitive process and an example, visual search process



neurological processes are involved with behavioural processes and phenotypes. NBO defines these relations but does not do so exhaustively. Since the number of classes in ANBO is small relative to NBO itself, we were able to define relationships for all cognitive processes included in the ontology, listed in Table 2. A second highlight to performing inferences is the proposed INPUT-OUTPUT

view of the cognitive process. This is always implemented in exactly the same way for every cognitive process. In any case, ANBO follows a similar approach to NBO in formalizing and making computationally accessible every cognitive process that belongs to ANBO (see Figure 2).

Relation	Definition	Example
in_response_to	Between a process x and a process y if and only if x occurs in response to y.	*auditory_behavior in_response_to hearing.* *visual_scanning in_response_to wakefulness*
by_means	A process x occurs by means of a material structure y if and only if x occurs by means of y.	*hearing behaviour by_means auditory system.* *visual_scanning by_means visual behaviour*
is_about	A process x is about some entity y if and only if x is about or directed toward y.	*shape constancy behavior is_about shape.* *visual search is_about physical object quality*
participates_in	A phenotype participates in a behaviour	*slow visual search phenotype* *participates_in some (has_quality some* *(increased duration and (towards some visual search)))*
has_input	A phenotype has input a collection of entities with a given property regarding frequency, amount and so on	

Table 2. Relations defined or used in NBO that are relevant for ANBO

4.1. Phenotype And Trait Ontology (PATO)

The Phenotype And Trait Ontology (PATO; Gkoutos et al., 2005, 2009) aims to capture information about phenotypes in any organism. This composes the phenotype by means of two variables: the entity that is observed and the specific characteristic or quality of that entity. Thus, PATO provides a framework for formalizing qualities which enables the characterization of every occurrence of a cognitive process. For example quality, *temporally extended* defined as "a quality of a process which ends later than the natural end time", could affect a *visual search* process that needs more time than usual. A second example is the quality *lacking processual parts* regarding *planning*. PATO defines this quality as "a quality of a process inhering in a bearer by virtue of the bearer's lacking a processual part as specified by the additional entity". The ANBO trait *planning* is a sequence of actions that will achieve a goal. Thus, ANBO makes use of *NBO:has_quality lacking procedural parts* property to express plannings where at least one step is missed.

4.2. Uberon

Uberon, is a cross-species ontology representing anatomical entities organized via anatomical classification reflecting morphology of organs and tissues (Mungall et al., 2012). It is species-agnostic, and includes annotations linking it to species specific (e.g., human or mouse) anatomical ontologies. Thus, it proves a bridge to facilitate cross-

species inference. In other words, Uberon provides several levels of detail for anatomical structures, mainly by means of *is-a* and *part-of* relations. The way that ANBO makes use of Uberon is twofold:

- In the same line as some processes defined in NBO, every ANBO cognitive process needs a given anatomical structure as a prerequisite. More concisely, the perception of a stimulus happens due to one or more sensory systems. That is the relation *by_means* draws in Figure 2.
- The user profile, which accesses users' ability to perform ADLs, is completed. As performance of ADLs is used to model cognitive function, and performance relies on a combination of cognitive and physical ability, this ensures that a user is screened to make sure that they are physiologically capable of performing the ADLs that will be monitored.

One advantage of using Uberon is that it is easy to increase the detail of sensory needs linked to a specific cognitive process. The current version of ANBO provides a quite high-level level of anatomical structures that are needed as part of the input of a cognitive process. For example, one user prerequisite that must be satisfied to accomplish *visual search* is that the user *visual system* must be functioning well enough to both visually scan and interpret input from this scanning. But future versions could distinguish in finer detail which part or parts of the visual system, such as *accessory optic system*[4], should be healthy instead of an all-or-nothing approach.

4.3. The Gene Ontology (GO)

GO (Ashburner et al., 2000; Consortium, 2004) provides structured, controlled vocabularies and classifications that cover several domains of molecular and cellular biology. The behavioral process branch of NBO contains a classification of behavior processes which complement and extend GO's behavior process domain. In the same way, ANBO both reuses and expands several classes related to perception, such as GO *visual perception*, *auditive perception* and ANBO *hand proprioception*, which extends GO *proprioception*.

5. A case of use: Ontology SmartLab Elderly (OSLE) and Telehealth Smart Home

The first example of an application of ANBO is to give support to the hypothesis that it is possible to infer cognitive decline by means of performance of daily activities. It is important to note that the objective of this section is not to provide evidence of the hypothesis directly but to validate ANBO by applying it in a specific domain, more concisely the Telehealth Smart Home (TSH). The rest of this section is structured as follows: we briefly introduce the domain where ANBO is applied, TSH. Then we describe Ontology SmartLab Elderly (OSLE), our own ontology integrating with TSH technology. Finally we show how ANBO and OSLE are integrated by means of a number of SWRL translation rules. SWRL is a rule language defined to complement OWL functionality (Horrocks et al., 2004).

[4] According to the Uberon definition, *accessory optic system* as "subdivision of visual system that processes movements of images across the retina and regulates the movement of eyes to keep the image stable"

5.1. Telehealth Smart Home and related studies

Telehealth Smart Home is defined as an adequate model of a smart home designed to care for someone with loss of cognitive autonomy (Rialle et al., 2002). Under this view, a model for taking care of an elderly person suffering loss of cognitive autonomy is proposed by Latfi et al. (2007). They propose seven ontologies that cover several relevant domains associated with the task such as *PersonAndMedicalHistory* or *BehaviourOntology*. Then a number of Bayesian networks that are part of the activity recognition are initiated by means of the instances of the ontologies. These Bayesian networks are used to recognize the activity the patient is probably performing. They are also involved in the learning process of the life habits of the Telehealth Smart Home occupant. As a consequence, they defined a a hierarchy of Bayesian networks which is closely related to the hierarchy of activities. At the lower level are specialized networks devoted to the recognition of simple activities such as the ones which can take place in the bathroom in front of the wash basin. The structure of these networks is defined using the instances of the corresponding ontologies.

In a similar way, Nachabe et al. (2016b,a) propose *OntoSmart* which is focused on monitoring elderly people's activity at home by means of a a wireless sensor network attached to both the body of the patient and sensors/actuators related to ambient parameters and home appliances. Then they implement rules in order to detect and notify atomic activities such as "the patient is standing" or abnormal values related to health alerts such as a heart attack or stroke.

Another example is AGNES (Peter et al., 2013), a system that monitors well-being (the person is happy or unhappy), activity (the person is physically active, very active or resting) and presence (the person is at home) by using several sensors and devices of common use such as smart watches, mobile phones, ambient displays and web cams. This information is then relayed to selected members of the social network of that person.

Hong et al. (2009) proposes an ontology for activity monitoring of daily living aimed at the elderly and disabled. The interaction with objects and movements involved in an activity are recorded by associated sensors which send signals to the central management system for processing. Every sensor signal is attached to a context that is part of an activity of interest. The interrelationships among sensors, contexts and activities is represented by a hierarchical network of ontologies where every node is a sensor, a context or an activity. In addition, it is possible to represent dependencies among nodes by means of AND and OR arcs. Similarly, it distinguishes compulsory, optional and compound nodes.

Finally, González-Landero et al. (2019) propose a mechanism of measuring memory with a "smart cupboard", a cupboard with three sensorized doors with magnetic door sensors. The main goal was to have a device able to assess the memory in a familiar environment without requiring additional effort from the user. For this end, three algorithms are proposed to determine when a user finds an item, and when the user searches for an item without success: the first case is that the user finds a certain item in the first attempt. The second case is that the user finds a certain item in a certain number of attempts. The last case is the one in which the user did not find the item.

5.2. Ontology SmartLab Elderly (OSLE)

OSLE is our proposal to model activities of daily living (ADL). OSLE explicitly represents both activities as a sequence of predefined steps and activities without a specific structure, just a sequence of actions that happens in a given time window. The key differece between both types of activities is that, for the first case, the system "knows" what is the objective and structure of the activity. For example, washing dirty laundry. This is a knowledge-based approach, and OSLE follows the work reported in Hong et al. (2009), briefly introduced previously. In this way, OSLE is made up of sensors, contexts that are the interpretation of sensor readings, and activities that are defined as a group of contexts and/or other activities. For the second case, the task that is accomplished by the user is not modeled as part of the ontology. Activities are not described, but the registration of the sensors and the order in which every sensor activation happens, although such activities are not necessarily part of a predefined process hard coded in the ontology. It is a data-driven approach such as is proposed in Salguero and Espinilla (2017), allowing flexible annotation to the ANBO/OSLE system.

5.2.1. Implementation of OSLE

OSLE is implemented as a specialization of Ontology for Biomedical Investigations, OBI (Bandrowski et al., 2016; Peters et al., 2009). OBI is part of the OBO Foundry which include NBO, GO and PATO. As a consequence, combining OSLE and ANBO is a relatively easy process (see section 5.2.4).

OSLE as an extension of OBI distinguishes between the specification of a plan (*obi: plan specificaction*) and the *realization* of that plan (*obi: planned process*) once this plan is *concretized*. As a specialization of these concepts, OSLE defines *osle:ADL specification* whose realization is achieved by means of ADL processes. At this point, OSLE defines a plan specification as a sequence of *osle:ADL specification step*s. In order to declare each step, a *osle:context* is required which is attached to a given *osle:sensor* and may be a position in the sequence of steps. Finally, a *osle:daily living action* is the register of a context that is triggered as a consequence of the activation of a sensor in a given time-stamp. Eventually, a *osle:daily living action* is attached to *osle:ADL specification step*. An overview of OSLE is depicted in Figure 3.

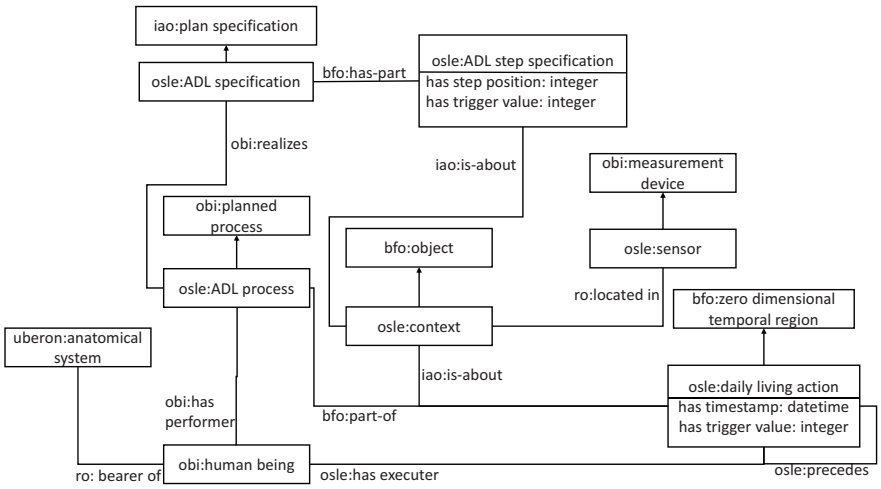
5.2.2. Creating activities in OSLE

In favour of greater clarity, we include the sequence of steps to be followed to both create a new daily living activity specification and register occurrences of such activity or just sequences of sensor readings (contexts) over the course of the day:

Defining a new type of activity

1. Create a new activity (*osle:ADL specification*). For example, *clean dirty clothes using the washing machine*
2. Create contexts as needed (*osle:context*). For example, *washing machine door*
3. Declare sensors as needed (*osle:sensor*) and attach them to the corresponding context (*obi:part of* property). For example, sensor *D09* is attached to *washing machine door*.

Figure 3. Ontology SmartLab Elderly class diagram.



4. Define the sequence of activity steps (*osle:ADL step specification*). Every step is the activation of a given context with a specific value related to an activity. Optionally, it is possible to include the expected order of the step. For example. step *a_cdc step 2* is the second step of activity *clean dirty clothes using the washing machine* and it defines that the value of this context must be *open* (the person opened the door of the washing machine).

Recording the occurrence of an activity

1. Declare the activity performer if it is not previously defined (mp:human being instances)
2. Makes concrete the activity to be performed (*bfo:specifically dependent continuant obi:concretizes osle:ADL specification*)
3. Declare a new *osle:ADL process* as the realization of the concretion of an activity
4. Record every daily living action performed by a *mp:human being* into a sequence of actions over the course of a period of time.

5.2.3. Instance of OSLE

At this moment, we have modeled a smart home equipped with 21 sensors (see Figure 4). All of them have been calibrated by repeating every action 50 times and noting down the number of false readings for each sensor. Since these sensors are quite straightforward to use, we have found no fails in our readings. By means of these sensors we have defined 15 different ADLs: *Clean dirty clothes, Take prescribed medicine, Make breakfast, Make lunch, Make dinner, Have a snack Watch television, Go home, Play a video game, Brush your teeth, Use the toilet, Do the dishes, Change clothes* and *Go to bed*. It is important to remark that the smart home depicted in Figure 4 is an artificial environ-

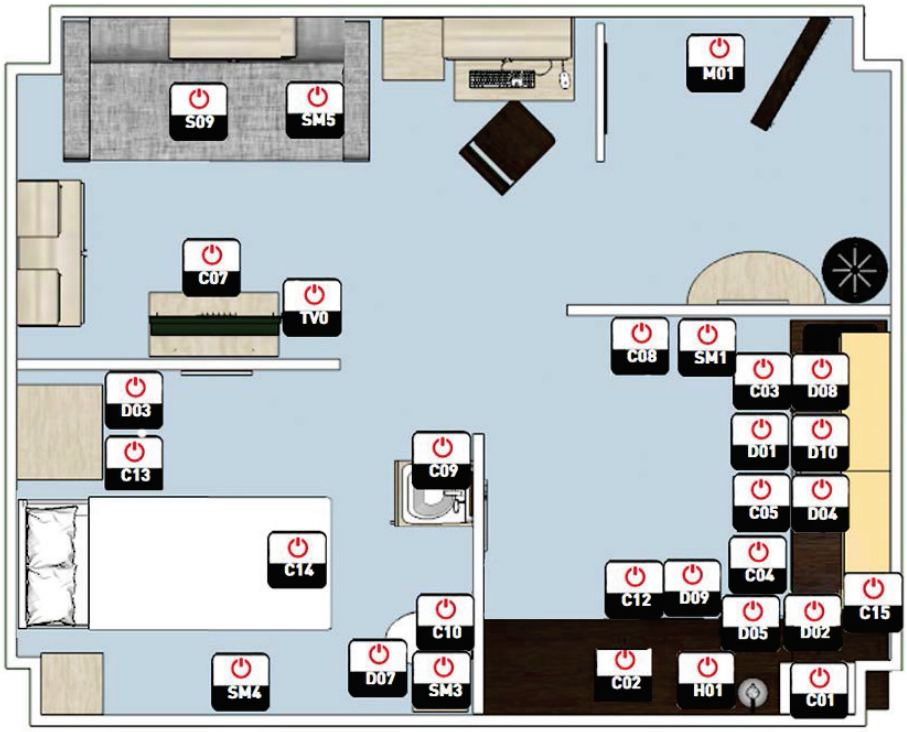


Figure 4. Sensors map in SmartLab: door(M01),TV(TV0), kitchen movement(SM1), motion bathroom(SM3), motion bedroom(SM4), motion sofa(SM5), refrigerator(D01), microwave(D02), wardrobe clothes(D03), cupboards cups(D04), dishwasher(D05), top WC(D07), closet(D08), washing machine(D09), pantry(D10), kettle(H01), medication box(C01), fruit platter(C02), cutlery(C03), pots(C04), water bottle(C05), remote XBOX(C07), trash(C08), tap(C09), tank(C10), laundry basket(C12), pyjamas drawer(C13), bed(C14), kitchen faucet(C15), pressure sofa(S09)

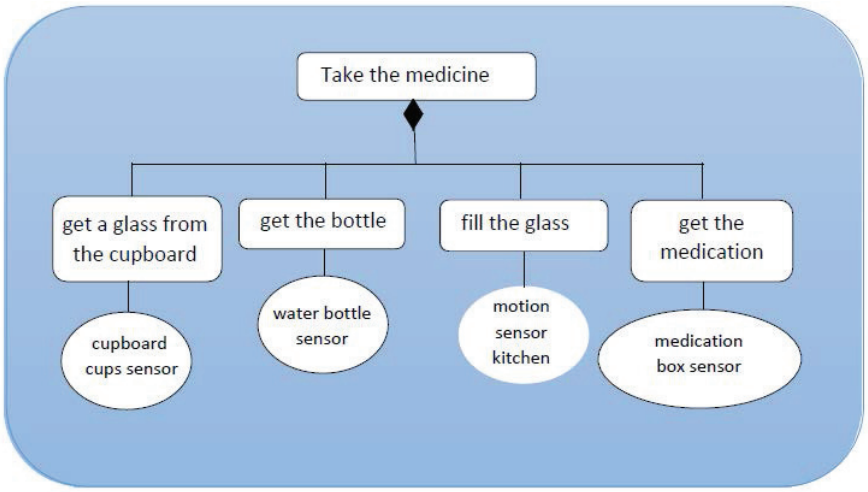
ment that is suitable in order to test the proposed implementation for each ADL. In this way, we have instantiated the proposed ADLs in order to check that, given an ADL, (i) the ADL is correctly identified and (ii) there is not performance issues. For this end, the current prototype has been executed in the main node of a cluster[5] that is made up by 2x Intel(R) Xeon(R) Silver 4210 CPU @ 2.20GHz processors, 96 GB DDR4 @ 2933 Mhz of memory and 60TB of storage. We have checked that ADLs are recorded in real time with no noticeable delay.

5.2.4. Integration of OSLE and ANBO

The end of the integration of OSLE and ANBO is to grasp the relationship between ADLs and cognitive decline associated with the elderly. For example, cooking a french omelette implies the activation of a number of cognitive processes such as is depicted in Figure 6. Even though neither our smart home installation nor OSLE provide the level of detail needed to detect every step of this example, it highlights the potential of integrating

[5]The complete description of the hardware is available at https://www.ujaen.es/centros/ceatic/servicios/supercomputacion/cluster-ada [11-01-2020]

Figure 5. Example of an activity. Notation is following Yao et al. (2005): squares are contexts and activities, ellipses are sensors and the diamond is an AND connector

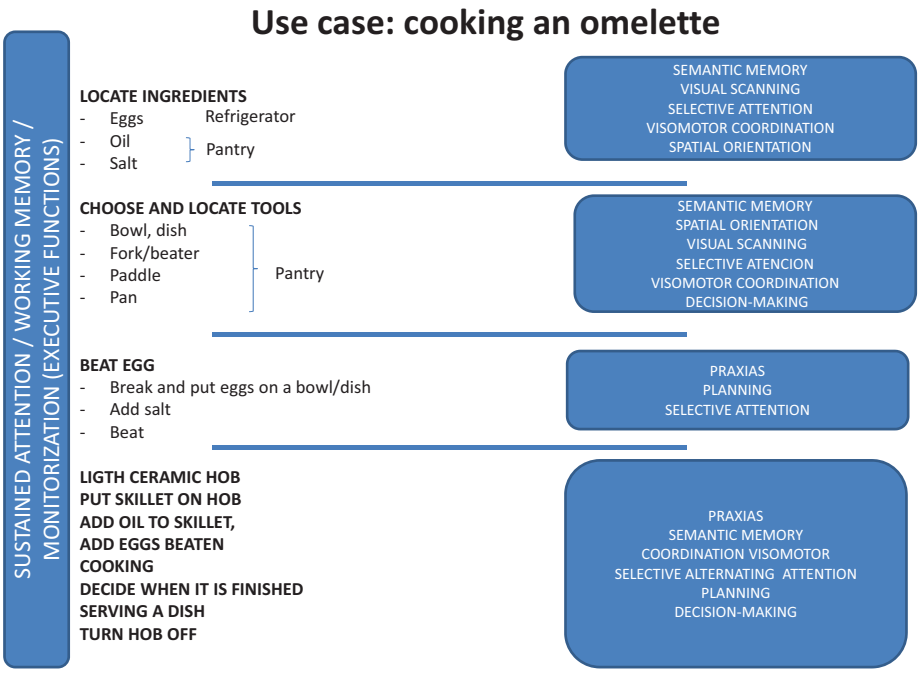


ANBO and OSLE. Our working hypothesis is that more simple or general activities such as choosing the correct medication in the medicine box or cooking are suitable for detecting some clues or issues regarding cognitive processing. The activities that are currently monitored in SmartLab are shown in Figure 4. For example, activity *take a medicine* defines that is necessary to fetch the correct medicine from the medication box. Whether the correct medicine is chosen or not is a clue about the performance of visual search or working memory processes. Therefore, the aim is to give support to this type of reasoning by activating ANBO concepts as a consequence of events or qualities of both environment and user profile registered in OSLE. From this point of view, the integration of ANBO and OSLE is a matter of interpretation of OSLE events as ANBO processes. This approach is implemented as follows:

- The OSLE source is sensor readings. Sensor readings are attached to a given physical context with some sensory qualities. For example, sensor *C05* is attached to a water bottle so when the sensor is activated, it is interpreted as the inhabitant obtained the bottle. In addition, this action has some physical qualities that are suitable for perception by sensory systems such as visual or touch systems.
- Within the ANBO source, every cognitive process requires some prerequisite input layer to be activated. For example, *visual search* (i) is triggered *in response to a visual perception by means of a visual system* of *physical object qualities* and (ii) *is about a physical object quality*, such as morphology or colour of the object or group of objects scanned by means of the vision system.
- The act of obtaining the water bottle is interpreted as a process of type *visual search*, inter alia. In order to accomplish this kind of translation or interpretation, a number of SWRL rules are applied. For the example given, some examples of the rules that are applied are the following (see Figure 7):

```
1.
Name: LightQualityToPhysicalObjectQuality
Comment: Translates a OSLE  context with a visual attribute (color, shape...)
```

Figure 6. Cooking a omelette illustrates the kind of relationship that it is possible to define between both domains, ADL and cognitive functions. Nevertheless, this is a quite ambitious example from the point of view of sensors that are needed to register every action and it exceeds the the current capabilities of our SmartLab

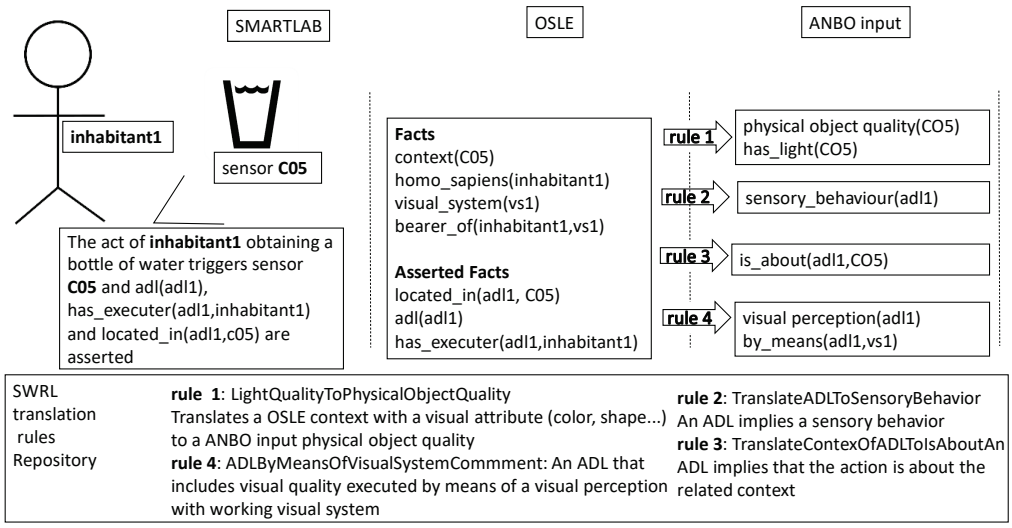



```
        to a ANBO input physical object quality
  OSLE_0000103: context
  PATO_0001241: physical object quality
Rule:
obo1:OSLE_0000103(?c)
^ obo1:uberon/has_quality(?c, osle:light)
->  obo1:PATO_0001241(?c)

2.
Name: TranslateADLToSensoryBehavior
Commment: An ADL implies a sensory behavior
Required to infer: visual behavior
OSLE_0000110: daily living action
  NBO_0000308: sensory behavior
Rule:
obo1:OSLE_0000110(?d)
->obo1:NBO_0000308(?d)
```

```
3.
Name: TranslateContexOfADLToIsAbout
Commment: An ADL implies that the action is about the related context
  OSLE_0000110: daily living action
  RO_0001025: located in
OSLE_0000103: context
Rule:
obo1:OSLE_0000110(?d)
^ obo1:RO_0001025(?d,?c)
^ obo1:OSLE_0000103(?c)
-> nbo:is_about(?d, ?c)
```

Figure 7. An example of SWRL application: an inhabitant gets a bottle of water and it is interpreted as an ADL that is part of the input of a visual search process.



By means of of these rules and others like them, ADLs are interpreted in an incremental way. For example, rule 1 asserts that a glass of water has visual properties, no matter if the attached sensor is triggered or not. Once sensor activation happens, this is interpreted as a visual stimulus. In the case of the elderly person owning a working visual system then the visual stimulus is perceived by means of such a visual system (rule 3). Finally, the combination of rules 1 and 4 allows us to make the inference that a visual search process occurs.

Inferences such as the one depicted above make it possible to ensure independence between the domains concerned: smart homes and cognitive processes. Thus, if a new ADL is defined, it does not include any kind of information about the potential cognitive processes that are implied, just the physical and logical properties of the activity and objects that are part of the new ADL. In the same way, if a new cognitive process is

integrated into ANBO, it is defined in terms of input and output parameters, that are attached to an ADL by applying the general SWRL rules.

Moreover, complex tasks that imply several steps allow us to detect divergences between the expected result and the real result, so these divergences could be interpreted as a fail or malfunction of a cognitive process and, in the end, a sign of cognitive decline. For example *osle:take the medicine* requires *get the medication*, *get a glass* and *filling* it. Suppose that the step *filling a glass* never happens, then it is annotated as a fail that is related to cognitive processes such as *anbo:planification* or *ambo:short term memory*. Of course, it is not possible to infer a cognitive decline just because of this divergence between the specification of the ADL and a specific occurrence of the activity. But findings such as not filling the glass when expected are noted down so tendencies of user behavior can be detected by the expert. In other words, the dual ANBO-OSLE is able to comply with requests such as to retrieve a "list of failed ADLs related to planification".

Relation	Definition	Example
part-of	a sensory system is part-of a human	visual system
is-about	an activity step is about a cognitive process	get the medication is about take the medicine

Table 3. ANBO and NBO relations

5.2.5. A note about scalability

Such as is explained above, the proposed architecture has been deployed in a only scenario, the SmartLab, with the main aim of testing and debugging the implementation of ADLs. As a consequence, this says nothing about the scalability of the system towards real environments where the activity of an undetermined number of users is recorded. In any case, the integration of ANBO with OSLE, and particularly the use of reasoning with SWRL, is suitably efficient as to be used in real systems. For this reason, the tractability of the system has been considered in several respects:

- Both ontologies are implemented by means of DL subset of OWL, a standard formalization in Description Logic, maintaining decidability. Moreover, The ELK reasoner can be used for guaranteed classification in polynomial time (Kazakov et al., 2014).
- The ontology design follows OBO Foundry (Smith et al., 2007) recommendations. That is, inter alia, non-overlapping, strictly-scoped content and imports following the MIREOT strategy. These ensure that the ontology is efficient, and that minimal fragments are included from external ontology.
- The possible tree of rules instantiated by the SWRL rules is low in depth and breadth, due to short chaining and a small set of sensors to be triggered together in a temporal period (as the user interacts with the home environment) respectively.

6. ANBO Evaluation

This section complements the validation accomplished by applying ANBO to TLH depicted in section 3, justifying the correctness of ANBO under the view of validation

carried out according to the measures proposed by Gómez-Pérez (2001) also together with Lopez and Corcho (2004): consistency, completion, conciseness, expandability and sensitiveness. The following lines justify that ANBO satisfies these criteria:

- Consistency: ANBO, OSLE, the SWRL rules, and all components in combination, are logically consistent. When evaluated using the Hermit reasoner (Shearer et al., 2008), which evaluates the consequences of all asserted axioms in an ontology, no cases of ontology inconsistency, nor class unsatisfiability (classes which logically cannot have instances) were detected.
- Completion: In brief, all that is supposed to be in the ontology is explicitly set out in it, or can be inferred. The way to check completion is by means of the integration with OSLE and the definition of at least one ADL for each cognitive process. Such as is depicted in section 5, OSLE is independent from ANBO and the SWRL are independent from the cognitive processes. As a consequence, every cognitive process is inferred rather than assessed. In addition, as intermediate steps over the course of these inferences, the reasoner obtains instances of ANBO concepts related to sensory stimulus and physical object qualities.
- Conciseness: Again, we rely on OSLE to check this desirable property of the ontology: every class that is explicitly defined in ANBO is instantiated as a consequence of at least one ADL. This proves that every concept is useful. In addition, we have explicitly defined as disjoint classes those that share a common superclass. For example, the whole of the cognitive processes of interest for ANBO are defined as disjoint classes.
- Expandability is related to "the effort required in adding new definitions to an ontology and more knowledge to its definitions, without altering the set of well-defined properties that are already guaranteed"(Gómez-Pérez, 2001). Any necessary expansion of the ontology is a task that is well-defined and homogeneous, since we have carefully defined a template, depicted in Figure 2, such that a new cognitive process is always defined in the same way: (i) extending NBO *cognitive behavior* class and (ii) defining the input and output layers always in terms of sensory and environmental requirement and outcomes. Since new definitions would rely on including and expanding existent NBO concepts, we ensure that descriptions of new cognitive processes can be developed with minimal overhead.
- Sensitiveness relates to how small changes in a definition alter the set of well-defined properties that are already guaranteed. We face this issue both when a cognitive process is added or modified and when ANBO as a whole is integrated with other ontologies such as OSLE. Cognitive processes are independent, and therefore do not require knowledge about other cognitive processes - this means that addition of new cognitive processes and modification of existing processes will not have knock-on effects outside of the relevant semantic locality. The use of SWRL rules as a bridging layer between OSLE and ANBO ensures that details that pertain to particular application uses are not included in the ontology itself, meaning that definitions will not need to be changed to support additional uses. Since the cognitive processes in ANBO make heavy use of well-developed and supported ontologies in the relevant domains, major changes are unlikely to be made, or will be made in such a way that support backwards-compatibility, and a preservation of logical functionality.

7. Conclusions and future work

Given an aging population and associated cognitive decline, identifying factors that may shield individuals from cognitive deterioration is a matter of societal interest. Many studies have attempted to identify ways to detect, reduce and/or counteract the course of cognitive and brain decline. However, traditional neuropsychological methods have some limitations, such as poor ecological validity. In this paper we present ANBO, a framework that builds on well established methodologies in knowledge representation, that can be used as a model of neuropsychological function with respect to ADLs, in order to create a method by which cognitive decline may be detected, and appropriate interventions determined. In addition, we implement and demonstrate this framework, by applying ANBO to Telehealth Smart Homes, allowing us to describe Activities of Daily Living. As a result of such integration, it is possible to make an original interpretation of these activities in terms of cognitive processes their success. The work in describing cognitive processes also has implications in the development of a axiomatisation pattern for cognitive processes, which could be used to normalize axioms in higher level concepts.

The work presented may be continued and extended in many ways. A point of importance in establishing the validity of this method is to run an actual trial involving test subjects, and using the generated data to determine whether neuropsychological decline can be adequately detected. NBO should also be applied to design interventions integrated as naturally as possible into ADLs in order to prevent and/or palliate cognitive decline. There are also plenty of opportunities to increase the coverage of ANBO (i) by including new cognitive processes of interest in the field of the elderly and (ii) by giving more detail. More detailed definitions of cognitive processes would make it possible to be more precise regarding the sensory capabilities required for a given cognitive process, thereby improving the sensitivity of any inferences. The system could also be applied, through development of other SWRL interaction layers, to other ontology-based smart-home systems (as mentioned earlier). In this way, from our point of view, one of the most promising ways to apply this ontology is the implementation of, i.e., a bayesian model in order to learn from the user behaviour so that, in the long term, it enables a kind of cognitive decline alert system when deviation related with the user's behaviour are detected and this deviations are statiscally significant. From this point a view, a isolated fail will no be relevant, but several of then could trigger the alert about a specific cognitive domain decline. It is also our intention to apply ANBO in other domains, such as in cognitive rehabilitation. In effect, the model would be applicable in any domain which seeks to associate physical behaviours with cognitive processes.

Regarding ANBO and OSLE integration, the next step is clear: it is necessary to validate the scalability of the architecture by monitoring a significant number of real homes. Note that it is difficult to replicate the whole SmartLab in every home. For this reason, both a minimum set of sensors and ADLs will be defined. At the moment of writing this paper, task "watch TV" is ready to be implemented in real homes by means of cheap sensors (TV0,C07,So9 and SM5 sensors such as are depicted in Figure 4), and a Raspberry Pi 3b+ plus a 4G communication module (Waveshare Hat SIM7600E) as base gatway. As a second stage will be the validation of the hypothesis that it is possible to identify cognitive decline, more concisely memory decline, by means of the integration of both OSLE and ANBO.

A. URLs of ANBO, OSLE and related ontologies and resources

In this appendix, we list the URLs where the referenced ontologies described in section 3 and 5.2 are available.

Related ontology	Type	Source	Definition
ANBO	Ontology	http://www4.ujaen.es/ dofer/ontologies/ obo/releases/2018-31-07/anbo.owl	Formal model of cognitive processes that are particularly relevant on a day-to-day ba-sis and whose performance usually declines when adults get older
OSLE	Ontology	http://www4.ujaen.es/ dofer/ontologies/ obo/releases/2018-31-07/osle.owl	An ontology related with Telehealth Smart Homes where Activities of Daily Living are recorded
OSLE example	Ontology	http://www4.ujaen.es/ dofer/ontologies/ obo/releases/2018-31-07/ osle_instance1.owl	An instance of OSLE according Smart Lab configuration (see Figure 4 and 5)
ANBO & OSLE	SWRL	http://www4.ujaen.es/ dofer/ontologies/ obo/releases/2018-31-07/trules.swrl	Rules in order to integrate ANBO and OSLE while keeping both ontologies independet
ANBO	imports	http://www4.ujaen.es/ dofer/ontologies/ obo/anbo/external/nbo_import.owl http://www4.ujaen.es/ dofer/ontologies/ obo/anbo/external/nbo_import_mireot.owl	ANBO imports from NBO and related ontologies
ANBO	OntoFox config file	http://www4.ujaen.es/ dofer/ontologies/ obo/anbo/external/nbo_import_mireot.txt	Ontofox input file in order to generate nbo_import_mireot.owl imports
OSLE	imports	http://www4.ujaen.es/ dofer/ontologies/ obo/osle/external/go_import.owl http://www4.ujaen.es/ dofer/ontologies/ obo/osle/external/obi_import.owl http://www4.ujaen.es/ dofer/ontologies/ obo/osle/external/uberon_import.owl	OSLE imports from OBI and related ontologies

References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25.

Atkinson, R. C. and Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes1. In *Psychology of learning and motivation*, volume 2, pages 89–195. Elsevier.

Baddeley, A. (1992). Working memory. *Science*, 255(5044):556–559.

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, 4(11):417–423.

Ball, K., Berch, D. B., Helmers, K. F., Jobe, J. B., Leveck, M. D., Marsiske, M., Morris, J. N., Rebok, G. W., Smith, D. M., Tennstedt, S. L., et al. (2002). Effects of cognitive training interventions with older adults: a randomized controlled trial. *Jama*, 288(18):2271–2281.

Ballesteros, S., Mayas, J., and Manuel Reales, J. (2013). Does a physically active lifestyle attenuate decline in all cognitive functions in old age? *Current aging science*, 6(2):189–198.

Bandrowski, A., Brinkman, R., Brochhausen, M., Brush, M. H., Bug, B., Chibucos, M. C., Clancy, K., Courtot, M., Derom, D., Dumontier, M., et al. (2016). The ontology for biomedical investigations. *PloS one*, 11(4):e0154556.

Brown, P. J., Devanand, D., Liu, X., and Caccappolo, E. (2011). Functional impairment in elderly patients with mild cognitive impairment and mild alzheimer disease. *Archives of general psychiatry*, 68(6):617–626.

Bruce, V. and Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, 77:305–327.

Consortium, G. O. (2004). The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(suppl_1):D258–D261.

Courtot, M., Gibson, F., Lister, A. L., Malone, J., Schober, D., Brinkman, R. R., and Ruttenberg, A. (2011). Mireot: The minimum information to reference an external ontology term. *Applied Ontology*, 6(1):23–33.

Engvig, A., Fjell, A. M., Westlye, L. T., Moberget, T., Sundseth, Ø., Larsen, V. A., and Walhovd, K. B. (2010). Effects of memory training on cortical thickness in the elderly. *Neuroimage*, 52(4):1667–1676.

Farias, S. T., Mungas, D., Reed, B. R., Cahn-Weiner, D., Jagust, W., Baynes, K., and DeCarli, C. (2008). The measurement of everyday cognition (ecog): scale development and psychometric properties. *Neuropsychology*, 22(4):531.

García-Viedma, M. R. and Fernández-Guinea, S. (2010). Neuropsicolog'ıa cognitiva: procesos psicológicos básicos. *Terapia ocupacional aplicada al daño cerebral adquirido*, pages 93–106.

Gkoutos, G. V., Green, E. C., Mallon, A.-M., Hancock, J. M., and Davidson, D. (2005). Using ontologies to describe mouse phenotypes. *Genome biology*, 6(1):R8.

Gkoutos, G. V., Mungall, C., Dolken, S., Ashburner, M., Lewis, S., Hancock, J., Schofield, P., Kohler, S., and Robinson, P. N. (2009). Entity/quality-based logical definitions for the human skeletal phenome using pato. In *Annual International Conference of the IEEE*, pages 7069–7072. Engineering in Medicine and Biology Society.

Gkoutos, G. V., Schofield, P. N., and Hoehndorf, R. (2012). The neurobehavior ontology: an ontology for annotation and integration of behavior and behavioral phenotypes. In *International review of neurobiology*, volume 103, pages 69–87. Elsevier.

Gómez-Pérez, A. (2001). Evaluation of ontologies. *International Journal of intelligent systems*, 16(3):391–409.

González-Landero, F., García-Magariño, I., Amariglio, R., and Lacuesta, R. (2019). Smart cupboard for assessing memory in home environment. *Sensors*, 19(11):2552.

Hong, X., Nugent, C., Mulvenna, M., McClean, S., Scotney, B., and Devlin, S. (2009). Evidential fusion of sensor data for activity recognition in smart homes. *Pervasive and Mobile Computing*, 5(3):236–252.

Horrocks, I., Patel-Schneider, P. F., Boley, H., Tabet, S., Grosof, B., Dean, M., et al. (2004). Swrl: A semantic web rule language combining owl and ruleml. *W3C Member submission*, 21(79).

Humphreys, G. W. and Riddoch, M. (2017). *Visual object processing: A cognitive neuropsychological approach, 1st ed.* Taylor & Francis Group.

Jekel, K., Damian, M., Wattmo, C., Hausner, L., Bullock, R., Connelly, P. J., Dubois, B., Eriksdotter, M., Ewers, M., Graessel, E., et al. (2015). Mild cognitive impairment and deficits in instrumental activities of daily living: a systematic review. *Alzheimer's research & therapy*, 7(1):17.

Karbach, J. and Kray, J. (2009). How useful is executive control training? age differences in near and far transfer of task-switching training. *Developmental science*, 12(6):978–990.

Kazakov, Y., Krötzsch, M., and Simančík, F. (2014). The incredible elk. *Journal of automated reasoning*, 53(1):1–61.

Latfi, F., Lefebvre, B., and Descheneaux, C. (2007). Ontology-based management of the telehealth smart home, dedicated to elderly in loss of cognitive autonomy. *OWLED*, 258.

Lawton, M. and Brody, E. M. (1970). Assessment of older people: self-maintaining and instrumental activities of daily living. *Nursing Research*, 19(3):278.

Lezak, M. D. (1995). *Neuropsychological assessment, 3rd ed.* Oxford University Press, USA.

Mcalister, C., Schmitter-Edgecombe, M., and Lamb, R. (2016). Examination of variables that may affect the relationship between cognition and functional status in individuals with mild cognitive impairment: A meta-analysis. *Archives of Clinical Neuropsychology*, 31(2):123–147.

Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E., and Haendel, M. A. (2012). Uberon, an integrative multi-species anatomy ontology. *Genome biology*, 13(1):R5.

Nachabe, L., Elhassan, B., Khawaja, J., and Salloum, H. (2016a). Semantic smart home system: ontosmart to monitor and assist habitant. *Int. J. Comput. Commun*, 10:78–86.

Nachabe, L., Girod-Genet, M., ElHassan, B., and Khawaja, J. (2016b). Ontology based tele-health smart home care system: ontosmart to monitor elderly. *Computer Science & Information Technology (CS & IT)*, pages 43–59.

Park, D. C., Lautenschlager, G., Hedden, T., Davidson, N. S., Smith, A. D., and Smith, P. K. (2002). Models of visuospatial and verbal memory across the adult life span. *Psychology and aging*, 17(2):299.

Pascual-Leone, A., Amedi, A., Fregni, F., and Merabet, L. B. (2005). The plastic human brain cortex. *Annu. Rev. Neurosci.*, 28:377–401.

Patterson, K. and Shewell, C. (2013). Speak and spell: Dissociations and word-class effects. *The Cognitive Neuropsychology of Language. Lawrence Erlbaum Associates, Inc*, pages 273–294.

Peter, C., Kreiner, A., Schröter, M., Kim, H., Bieber, G., Öhberg, F., Hoshi, K., Waterworth, E. L., Waterworth, J., and Ballesteros, S. (2013). Agnes: Connecting people in a multimodal way. *Journal on multimodal user interfaces*, 7(3):229–245.

Peters, B., Consortium, O., et al. (2009). Ontology for biomedical investigations. *Nature Publishing Group*.

Posner, M. I. and Dehaene, S. (1994). Attentional networks. *Trends in neurosciences*, 17(2):75–79.

Pugnetti, L., Mendozzi, L., Attree, E. A., Barbieri, E., Brooks, B. M., Cazzullo, C. L., Motta, A., and Rose, F. D. (1998). Probing memory and executive functions with virtual reality: Past and present studies. *CyberPsychology & Behavior*, 1(2):151–161.

Rialle, V., Duchene, F., Noury, N., Bajolle, L., and Demongeot, J. (2002). Health" smart" home: information technology for patients at home. *Telemedicine Journal and E-Health*, 8(4):395–409.

Rizzo, A. A., Schultheis, M., Kerns, K. A., and Mateer, C. (2004). Analysis of assets for virtual reality applications in neuropsychology. *Neuropsychological rehabilitation*, 14(1-2):207–239.

Rönnlund, M., Nyberg, L., Bäckman, L., and Nilsson, L.-G. (2005). Stability, growth, and decline in adult life span development of declarative memory: cross-sectional and longitudinal data from a population-based study. *Psychology and aging*, 20(1):3.

Salguero, A. and Espinilla, M. (2017). Improving activity classification using ontologies to expand features in smart environments. In *International Conference on Ubiquitous Computing and Ambient Intelligence*, pages 381–393. Springer.

Salthouse, T. A. (2010). Selective review of cognitive aging. *Journal of the International neuropsychological Society*, 16(5):754–760.

Shearer, R., Motik, B., and Horrocks, I. (2008). Hermit: A highly-efficient owl reasoner. *OWLED*, 432:91.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., et al. (2007). The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251.

Tulving, E. and Craik, F. I. M. (2000). The oxford handbook of memory. *Oxford University Press*.

United Nations, D. o. E. and Social Affairs, P. D. (2017). *World Population Aging. Report*. United Nations.

Valladares-Rodríguez, S., Pérez-Rodríguez, R., Anido-Rifón, L., and Fernández-Iglesias, M. (2016). Trends on the application of serious games to neuropsychological evaluation: a scoping review. *Journal of biomedical informatics*, 64:296–319.

Verghese, J., Lipton, R. B., Katz, M. J., Hall, C. B., Derby, C. A., Kuslansky, G., Ambrose, A. F., Sliwinski, M., and Buschke, H. (2003). Leisure activities and the risk of dementia in the elderly. *New England Journal of Medicine*, 348(25):2508–2516.

Wiggs, C. L., Weisberg, J., and Martin, A. (2006). Repetition priming across the adult lifespan—the long and short of it. *Aging, Neuropsychology, and Cognition*, 13(3-4):308–325.

Willis, S. L., Tennstedt, S. L., Marsiske, M., Ball, K., Elias, J., Koepke, K. M., Morris, J. N., Rebok, G. W., Unverzagt, F. W., Stoddard, A. M., et al. (2006). Long-term effects of cognitive training on everyday functional outcomes in older adults. *Jama*, 296(23):2805–2814.

Yao, H., Orme, A. M., and Etzkorn, L. (2005). Cohesion metrics for ontology design and application. *Journal of Computer science*, 1(1):107–113.

Young, A. W. and Bruce, V. (2011). Understanding person perception. *British Journal of Psychology*, 102(4):959–74.

DOI: 10.1111/dme.13870

Research: Care Delivery

Utility of HbA_{1c} assessment in people with diabetes awaiting liver transplantation

D. Bhattacharjee[1], S. Vracar[1], R. A. Round[2,3,4], P. G. Nightingale[4], J. A. Williams[4,5,6], G. V. Gkoutos[4,5,7,8,9,10], I. M. Stratton[11], R. Parker[12], S. D. Luzio[3,13], J. Webber[3], S. E. Manley[3,4,14], G. A. Roberts[3,13,15] and S. Ghosh[3,4]

[1]Medical School, University of Birmingham, Birmingham, [2]Clinical Laboratory Services, University Hospitals Birmingham NHS Foundation Trust, [3]Diabetes Translational Research Group, Diabetes Centre, Queen Elizabeth Hospital Birmingham, University Hospitals Birmingham NHS Foundation Trust, [4]Institute of Translational Medicine, University Hospitals Birmingham NHS Foundation Trust, [5]College of Medical and Dental Sciences, Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, [6]Mammalian Genetics Unit, Medical Research Council Harwell Institute, Harwell, [7]MRC Health Data Research UK (HDR UK), [8]NIHR Experimental Cancer Medicine Centre, Birmingham, [9]NIHR Surgical Reconstruction and Microbiology Research Centre, Birmingham, [10]NIHR Biomedical Research Centre, Birmingham, [11]Gloucestershire Retinal Research Group, Gloucestershire Hospitals NHS Foundation Trust, Cheltenham, [12]Leeds Liver Unit, St James's University Hospital, Leeds, [13]Diabetes Research Group, Swansea University, Swansea, [14]College of Medical and Dental Sciences, Institute of Metabolism and Systems Research, University of Birmingham, Birmingham and [15]HRB-Clinical Research Facility - Cork, University College Cork, Cork, Ireland

Accepted 22 November 2018

Abstract

Aims To investigate the relationship between HbA_{1c} and glucose in people with co-existing liver disease and diabetes awaiting transplant, and in those with diabetes but no liver disease.

Methods HbA_{1c} and random plasma glucose data were collected for 125 people with diabetes without liver disease and for 29 people awaiting liver transplant with diabetes and cirrhosis. Cirrhosis was caused by non-alcoholic fatty liver disease, hepatitis C, alcoholic liver disease, hereditary haemochromatosis, polycystic liver/kidneys, cryptogenic/non-cirrhotic portal hypertension and α-1-antitrypsin-related disease.

Results The median (interquartile range) age of the diabetes with cirrhosis group was 55 (49–63) years compared to 60 (50–71) years (P=0.13) in the group without cirrhosis. In the diabetes with cirrhosis group there were 21 men (72%) compared with 86 men (69%) in the group with diabetes and no cirrhosis (P=0.82). Of the group with diabetes and cirrhosis, 27 people (93%) were of white European ethnicity, two (7%) were South Asian and none was of Afro-Caribbean/other ethnicity compared with 94 (75%), 16 (13%), 10 (8%)/5 (4%), respectively, in the group with diabetes and no cirrhosis (P=0.20). Median (interquartile range) HbA_{1c} was 41 (32–56) mmol/mol [5.9 (5.1–7.3)%] vs 61 (52–70) mmol/mol [7.7 (6.9–8.6)%] (P<0.001), respectively, in the diabetes with cirrhosis group vs the diabetes without cirrhosis group. The glucose concentrations were 8.4 (7.0–11.2) mmol/l vs 7.3 (5.2–11.5) mmol/l (P=0.17). HbA_{1c} was depressed by 20 mmol/mol (1.8%; P<0.001) in 28 participants with cirrhosis but elevated by 28 mmol/mol (2.6%) in the participant with α-1-antitrypsin disorder. Those with cirrhosis and depressed HbA_{1c} had fewer larger erythrocytes, and higher red cell distribution width and reticulocyte count. This was reflected in the positive association of glucose with mean cell volume (r=0.39) and haemoglobin level (r=0.49) and the negative association for HbA_{1c} (r=–0.28 and r=–0.26, respectively) in the diabetes group with cirrhosis.

Conclusion HbA_{1c} is not an appropriate test for blood glucose in people with cirrhosis and diabetes awaiting transplant as it reflects altered erythrocyte presentation.

Diabet. Med. 36, 1444–1452 (2019)

Introduction

Diabetes is a leading cause of liver disease, with cirrhosis responsible for a considerable number of deaths in people with diabetes in the USA [1]. The association is mediated by multiple mechanisms including dyslipidaemia and altered hepatic fatty acid processing [2]. Peripheral insulin resistance may contribute to the development of diabetes in people with hepatitis

Correspondence to: Dr Susan Manley. E-mail: susan.manley@uhb.nhs.uk
D.B. and S.V. are equal first authors.
S.E.M., G.A.R. and S.G. are equal last authors.

What's new?

- HbA_{1c} may not be an accurate reflection of blood glucose for the diagnosis/monitoring of diabetes in people with other illnesses or on certain drugs; people with diabetes and liver disease awaiting transplantation are one such group.
- HbA_{1c} was found to be depressed relative to random plasma glucose by 20 mmol/mol in people with diabetes and cirrhosis (n = 28) compared to people with diabetes but no liver disease (n = 125); however, HbA_{1c} was elevated in one person with cirrhosis attributable to α-1-antitrypsin disorder.
- Compromised HbA_{1c} may be related to haematological differences associated with liver disease involving erythrocyte half-life, with shorter/longer times giving less/more opportunity for glycation of haemoglobin.

C [3] and cirrhosis [4]. Post-transplant diabetes is well recognized, with HbA_{1c} testing not being appropriate immediately afterwards as a result of post-transplant anaemia [5] and also rendered inaccurate by some drugs such as ribavirin which is used for hepatitis C treatment [6].

In 2011, the WHO introduced HbA_{1c} assessment for the diagnosis of diabetes mellitus [7]. HbA_{1c} is now widely used for this purpose in primary care, resulting in a doubling of the number of HbA_{1c} assessments requested, and a corresponding decrease in glucose measurement [8]. Since 2014, the use of HbA_{1c} testing has been included in the American Diabetes Association guidelines for the diagnosis of diabetes in hospital [9]. This recommendation has been confirmed by assessment of undiagnosed diabetes in white European people admitted to an Irish hospital [10]. However, whilst the WHO bulletin lists medical conditions and drugs that may affect HbA_{1c}, it provides no references to quantitative evidence [7].

Our hospital laboratory has reviewed HbA_{1c} test results, referring values below the reference range or very high values in people without a previous diagnosis of diabetes for urgent medical attention [8,11]. Evidence is accumulating that various co-existing conditions affect HbA_{1c} and result in misdiagnosis or mismanagement of diabetes [12]. Recently, HbA_{1c} was measured in 200 people with decompensated cirrhosis referred for liver transplantation. Measured HbA_{1c} values were significantly lower when compared with HbA_{1c} calculated from three previous glucose values [13].

Given these concerns, we investigated random plasma glucose and HbA_{1c} in people recruited for research into the relationships between glycaemic markers when attending diabetes clinics at the hospital, and in people with co-existing cirrhosis and diabetes awaiting liver transplant who had available data on glycaemic markers and Model for End Stage Liver Disease (MELD) scores [14].

Participants and methods

Ethics

The West Midlands Local Research Ethics Committee confirmed ethical approval for the Glucose Fructosamine and HbA_{1c} research study investigating the relationships between glycaemic markers in people attending the diabetes clinic at University Hospitals Birmingham NHS Foundation Trust. This study met the requirements of the current revision of the Declaration of Helsinki.

For people with diabetes attending liver clinics between June and September 2012, data were obtained from the electronic patient record for a registered, internal clinical audit (CAB-05641-13) at University Hospitals Birmingham NHS Foundation Trust.

Study cohort

The people with diabetes without liver cirrhosis included adults with no variant haemoglobin (n=125) who were recruited from the diabetes clinic at Queen Elizabeth Hospital Birmingham, University Hospitals Birmingham NHS Foundation Trust, UK, between June 2007 and June 2009.

The people with co-existing liver cirrhosis and diabetes comprised people from different parts of the UK with cirrhosis of the liver and diabetes, attending day clinics in the Liver Department at Queen Elizabeth Hospital Birmingham, UK, who were being considered for liver transplantation. Those attending between October 2008 and June 2012 were included in the clinical audit; in total, 240 people were reviewed. HbA_{1c} and random plasma glucose measurements, performed at Queen Elizabeth Hospital Birmingham laboratories, were available for 29 out of the 50 transplant candidates with both cirrhosis and diabetes. No other measure of glycaemic control was available to the study. Indications for transplant included one or more of the following complications of cirrhosis: spontaneous bacterial peritonitis; ascites; variceal bleed; and hepatic encephalopathy. Of the 29 participants in this cohort, 15 (52%) had non-alcoholic fatty liver disease, six (21%) had hepatitis C, three (10%) had alcoholic liver disease, two (7%) had hereditary haemochromatosis, one (3%) had polycystic liver and kidneys, one (3%) had α-1-antitrypsin-related liver disease and one (3%) had cryptogenic/non cirrhotic portal hypertension. Data were collected for this preliminary, clinical audit from the Birmingham Systems Prescribing Information and Communications System and the CDS Telepath Systems Ltd databases.

Measurements

All measurements were performed at University Hospitals Birmingham NHS Foundation Trust, with single measurements of HbA_{1c}, random plasma glucose, serum bilirubin and creatinine, and full blood count. Blood was collected

into fluoride oxalate vacutainers for glucose measurement. Biochemical variables were measured on Roche c8000 analysers (Roche Diagnostics Ltd, Burgess Hill, UK) and full blood count on Beckman DxH800 analysers (Beckman Coulter Ltd, High Wycombe, UK).

HbA$_{1c}$ was measured in EDTA blood using an International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) aligned TOSOH G8 ion exchange high performance liquid chromatography analysers (Tosoh, Reading, UK) before realignment of their calibrator downwards by the manufacturer in September 2013 [15]. People with abnormal haemoglobin were excluded because HbA$_{1c}$ is not reported by the laboratory in its presence, as were people with a total chromatogram area <500 as specified in the manufacturer's protocol for HbA$_{1c}$ measurement (~80g/l haemoglobin).

The MELD score was calculated using the formula: [0.957 × ln(serum creatinine) + 0.378 × ln(serum bilirubin) + 1.120 × ln(INR) + 0.643] × 10, with creatinine set to 4.06 for participants on haemodialysis [14]. The normal range for the MELD score is 0 to 6, with a score of 40 defined as gravely ill.

Statistical analysis

Data on participants without liver disease were entered into an Excel spreadsheet with robust quality assurance. Biochemical and haematological data were downloaded directly from the laboratory Telepath database. Clinical audit data for participants with diabetes and liver disease were accessed in the electronic patient record and entered into a pre-prepared Excel spreadsheet. Data were analysed with Microsoft Excel, Analyse-it Version 2.22 (Analyse-it Software Ltd, Leeds, UK), SPSS Statistics for Windows version 22.0 (IBM Corp., Armonk, NY, USA) and R version 3.4.0 [16].

The characteristics of the study cohort are presented in Table 1 as median and interquartile range (IQR), count or percentage, with Mann–Whitney or Fisher's exact tests used to compare the groups. Reference ranges were obtained from the hospital laboratory. Simple linear regression was used to assess the relationships between HbA$_{1c}$ and random plasma glucose, with Fig. 1 showing regression lines for both groups and 2SD lines for people with diabetes without cirrhosis. Residual analysis was performed to assess the fit of the model for the regression of HbA$_{1c}$ vs glucose. Some skewness in the HbA$_{1c}$ data was demonstrated in a Normal Q-Q plot of residuals, but there was no evidence of non-linearity. Log transformation of HbA$_{1c}$ values reduced the skewness, but did not affect the linearity, and yielded an R^2 value of 0.44 rather than 0.42. As both models are valid and give similar results, and given the ease of use of non-transformed data, we have not used the log transformation. This has the added advantage that the model is not dependent on the choice of units for HbA$_{1c}$.

Calculation of the difference in the HbA$_{1c}$ intercepts for the people with co-existing liver cirrhosis and diabetes, and people with diabetes without liver cirrhosis assumed the slopes were equal. The equality of the slopes was assessed by testing the glucose × group interaction term in a general linear model for HbA$_{1c}$, with glucose as a covariate and the group as a factor.

The correlation grid shows results for 27 people with diabetes and liver disease, and 123 with diabetes without liver disease (Fig. 2). Correlations for people with co-existing liver cirrhosis and diabetes are shown in the area of the grid above the diagonal and, for people with diabetes without liver cirrhosis, below the diagonal. The colour of the circles indicates whether the correlations are positive or negative. The intensity of the colour and the size of the circle are proportional to the correlation coefficients [17].

Pearson coefficients were calculated for pairwise groupings of each variable within the group, and displayed using the CORRPLOT package in the R program v. 0.84 [17]. Correlation coefficients were then compared using the R psych package v 1.7.8 [18]. Fisher transformations of correlation matrices were created to compare correlation coefficients within and between groups (psych::r.test function), and also when testing the independence of the two groups (psych::corrtest function).

Significance tests were performed by establishing the Z-score for the difference between the Fisher Z-transformed correlations when divided by the standard error of the difference between the two Z-scores. To confirm the assumption that the groups are two distinct populations, separable by the variables measured, a test of equivalence of the Fisher Z-score equivalents of the two correlation matrices was performed, which indicated two distinct groups (P<1.2e-06, Z-score of differences = 4.98).

The profile of the participant with α-1-antitrypsin-related liver disease was summarized graphically by expressing each value as a multiple of the median value for that variable in the group of people with diabetes without liver cirrhosis. The median values for the people with co-existing cirrhosis and diabetes disease (excluding the person with α-1-antitrypsin-related liver disease) were plotted similarly, Fig. 3.

Results

Characteristics of participants

There were no significant differences in age, gender or ethnicity, but serum creatinine was significantly lower in people with diabetes and cirrhosis (P=0.001, Table 1). Two distinct populations were identified when all the variables were considered (P<0.001).

Glucose and HbA$_{1c}$

Random plasma glucose concentrations did not differ, but HbA$_{1c}$ was substantially lower in people with liver disease:

Table 1 Characteristics of people with liver cirrhosis and diabetes awaiting transplant vs people with diabetes but no liver disease

	Reference range	People with cirrhosis and diabetes	People with diabetes and no liver disease	*P*
N		29	125	
Age, years		55 (49–63)	60 (50–71)	0.13
Men, *n* (%)		21 (72)	86 (69)	0.82
Ethnicity, *n*				
White European		27	94	0.20
South Asian		2	16	
Afro-Caribbean		0	10	
Other		0	5	
Severity of disease				
MELD score	<6	12 (9–17)*		
Creatinine†, μmol/l		77 (63–110)	98 (86–112)	0.001
Glycaemic markers				
Random plasma glucose, mmol/l		8.4 (7.2–11.2)	7.3 (5.3–11.5)	0.17
HbA_{1c}, mmol/mol	<48	41 (32–56)	61 (52–70)	<0.001
%	<6.5	5.9 (5.1–7.3)	7.7 (6.9–8.6)	
Haematology				
Red blood cell count, $\times 10^{12}$/l	Men: 4.2–5.7 Women: 3.8–5.1	3.6 (3.0–3.9)‡	4.7(4.3–5.0)	<0.001
Haemoglobin, g/l	Men: 135–180 Women: 115–165	106 (93–122)‡	137 (125–147)	<0.001
Haematocrit, l/l	Men: 0.40–0.54 Women: 0.37–0.47	0.32 (0.27–0.35)‡	0.40 (0.38–0.43)	<0.001
Mean cell volume, fl	80–99	91 (85–96)	86 (83–89)	0.001
Mean cell haemoglobin, pg	27–33	31 (28–33)	30 (28–31)	0.028
Mean cell haemoglobin concentration, g/l	315–365	339 (327–349)	341 (329–350)	0.473
Red cell distribution width, %	11–14	17 (15–18)*	13 (13–14)	<0.001
Reticulocyte count, $\times 10^9$/l	20–80	61 (47–71)	45(37–64)	0.005
Platelets, $\times 10^9$/l	150–450	103 (78–153)‡	251 (214–289)	<0.001
White cell count				
White cell count, 10^9/l	4.0–11.0	5.1 (4.3–6.8)	7.2 (6.2–8.8)	<0.001
Neutrophils, 10^9/l	2.0–7.5	3.4 (2.6–4.4)	4.3 (3.5–5.7)	0.002
Lymphocytes, $\times 10^9$/l	1.0–4.0	1.1 (0.7–1.3)	2.1 (1.8–2.5)	<0.001
Monocytes, $\times 10^9$/l	0.2–0.8	0.5 (0.4–0.6)	0.6 (0.4–0.7)	0.064
Eosinophils, $\times 10^9$/l	0.0–0.4	0.2 (0.1–0.3)	0.2 (0.1–0.3)	0.444

Median (IQR) interquartile range; otherwise *n* or %.
*Median higher than reference range. †Creatinine reference ranges dependent on age and gender. ‡Median lower than reference range.

median (IQR) 41 (32–56) mmol/mol [5.9 (5.1–7.3)%] vs 61 (52–70) mmol/mol [7.7(6.9–8.6)%]; (P<0.001, Table 1 and Fig. 1).

HbA_{1c} and glucose were positively correlated: r^2=0.34 in those with liver disease and r^2=0.30 in those without (P<0.001). Linear regression equations are cited in mmol/mol (IFCC) and % (Diabetes Control and Complications Trial/UK Prospective Diabetes Study) units:

liver disease: (mmol/mol) $HbA_{1c} = 3.0 \times RPG + 15.5$; or (%) $HbA_{1c} = 0.27 \times RPG + 3.6$

no liver disease: (mmol/mol) $HbA_{1c} = 1.8 \times RPG + 46.3$; or (%) $HbA_{1c} = 0.17 \times RPG + 6.4$,

where RPG is random plasma glucose. There was a significant difference of 20 mmol/mol (1.8%) HbA_{1c} (P<0.001) between the intercepts, assuming the slopes to be equal (P=0.12). A similar result was obtained when the data were restricted to white European people.

Haematology

There were major haematological differences between the groups, with fewer red blood cells, and lower haemoglobin and haematocrit levels in the group with diabetes and cirrhosis; with the median values lower than the reference ranges (Table 1). The red blood cell distribution width was higher in those with liver disease and above the reference range. The equivalent values for those with diabetes but no cirrhosis were within the reference ranges. Higher values for mean red blood cell volume and mean red blood cell haemoglobin were found in the group with diabetes and cirrhosis, indicating larger red blood cells, and also a higher reticulocyte count, indicating a shorter half-life. People with

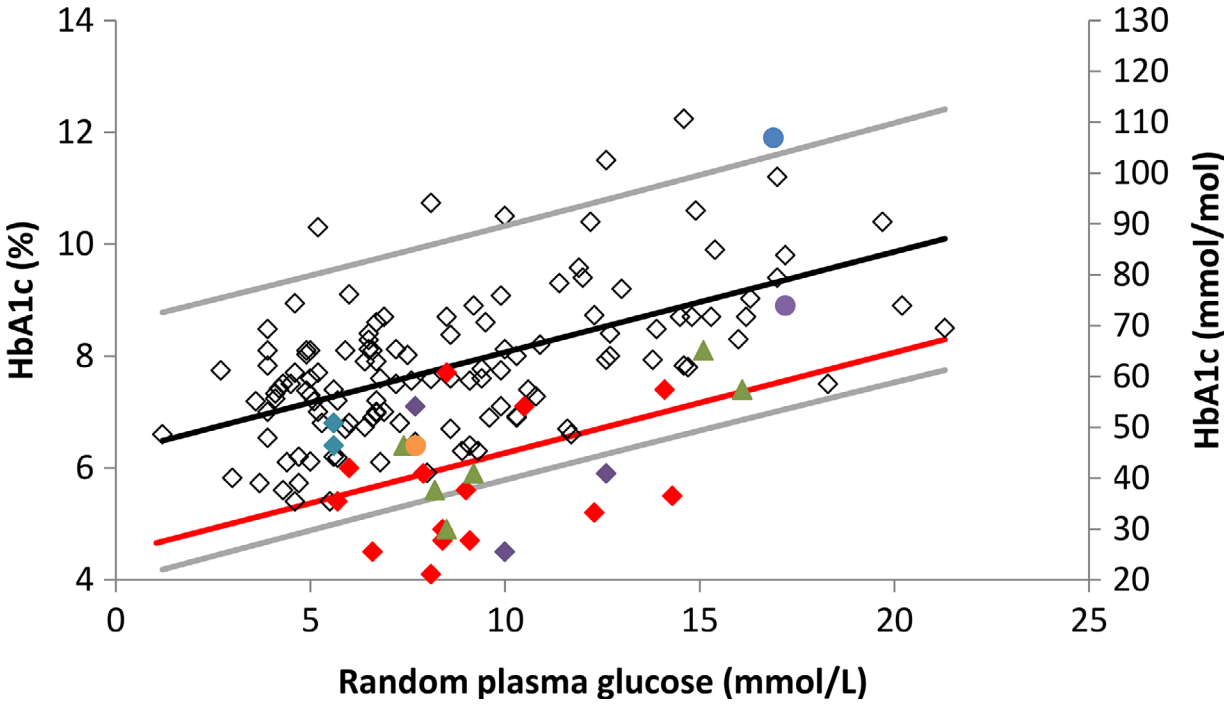


FIGURE 1 HbA$_{1c}$ and random plasma glucose in people with diabetes with cirrhosis awaiting liver transplant and without liver disease. The person with α-1-antitrypsin-related liver disease (blue circle) was excluded from further analyses. White diamonds: no liver disease; purple diamonds: alcoholic liver disease; red diamonds: non-alcoholic fatty liver disease; green triangles: hepatitis C; orange circle: polycystic liver and kidneys; blue diamonds: hereditary haemochromatosis; blue circle: α-1-antitrypsin-related liver disease; purple circle: cryptogenic/non-cirrhotic portal hypertension. Regression line black: no liver disease; regression line red: cirrhosis. 2SD lines: no liver disease.

diabetes and cirrhosis had fewer white blood cells, platelets and lymphocytes with no difference in eosinophils; all these counts were within the reference ranges in both groups.

Associations among variables

Further investigation was undertaken to determine the factors related to the depression of HbA$_{1c}$ in those with cirrhosis using a correlation grid (Fig. 2).

There were significant differences in the magnitude and direction of correlation coefficients for glucose between the groups: with mean cell haemoglobin: r=–0.092 (95% CI –0.265, 0.064) vs r=0.488 (95% CI 0.133, 0.732; P=0.010) in the diabetes without cirrhosis group vs the diabetes and cirrhosis group, respectively; mean cell haemoglobin concentration: r=–0.110 (95% CI –0.281, 0.069) vs r=0.363 (95% CI –0.018, 0.653), respectively (P=0.003, Fig. 2).

The correlation coefficients for HbA$_{1c}$ or glucose with the haematological variables showed statistically significant differences in the group with diabetes and cirrhosis. The correlation coefficients were positive for glucose, and negative or near zero for HbA$_{1c}$ for: (1) mean cell volume: HbA$_{1c}$, r=–0.278 (95% CI –0.595, 0.114); glucose, r=0.387 (95% CI 0.225, 0.528; P=0.020); (2) mean cell haemoglobin: HbA$_{1c}$, r=–0.260 (95% CI –0.583, 0.132); glucose, r = 0.488 (95% CI 0.341, 0.612; P = 0.010); (3) mean cell haemoglobin concentration: HbA$_{1c}$, r=–0.095 (95% CI –0.458, 0.296); glucose, r=0.363 (95% CI –0.018, 0.653); (P =0.049, Fig. 2).

When stepwise regression models were applied in those with diabetes and liver disease, red blood cell count and eosinophils had an R^2 value of 45.7% for HbA$_{1c}$, and mean cell haemoglobin and eosinophils an R^2 value of 39.9% for glucose. The most important factor determining HbA$_{1c}$ in people with diabetes but no liver disease was glucose.

Severity of liver disease

The median (interquartile range) Model for End Stage Liver Disease (MELD) score for the study cohort was calculated as 12 (9–17), (normal <6) in those with cirrhosis. The MELD score in people with co-existing liver cirrhosis and diabetes was negatively correlated with HbA$_{1c}$ (r=–0.56) and red blood cell count (r=–0.60). The correlation with glucose was r=–0.12, but MELD was positively correlated with mean cell haemoglobin, (r=0.32) and red cell distribution width (r=0.41).

α-1-antitrypsin disorder

The person with cirrhosis and diabetes related to α-1-antitrypsin disorder had high HbA$_{1c}$ relative to glucose, with HbA$_{1c}$ elevated by 28 mmol/mol (2.6%), (Fig 1). Their

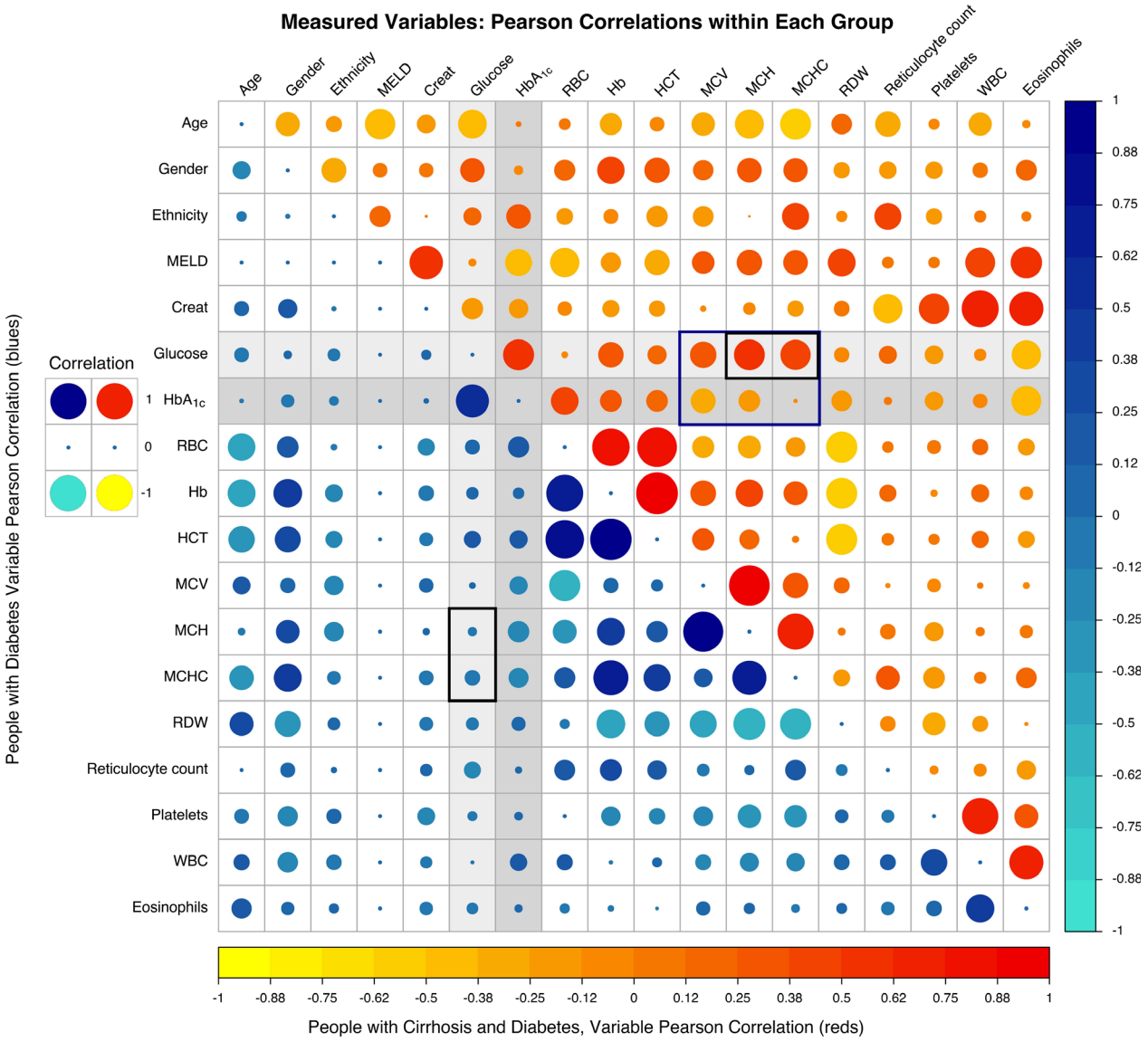


FIGURE 2 Relationships of characteristics of people with cirrhosis and diabetes awaiting liver transplant, and those with diabetes. Triangles: upper for those with liver disease, positive correlation coefficients, red and negative yellow; lower for those without liver disease, dark blue and cyan, accordingly. Circle size, largest for correlation +1 or –1; smallest if no correlation, i.e. 0. Shading: darker, HbA_{1c}; lighter, glucose. Box outlines: dark blue for statistically significant differences between HbA_{1c} and glucose correlations ($P<0.05$). Black for significant differences in correlations for glucose with variables. MELD, Model for End Stage Liver Disease; RBC, red blood cell count; Hb, haemoglobin; HCT, haematocrit; MCV, mean cell volume; MCH, mean cell haemoglobin; MCHC, mean cell haemoglobin concentration; RDW, red cell distribution width; WBC, white blood cell count.

haematological profile was different from that of the other people with cirrhosis and those without cirrhosis. The plot of haematological data for the person with α-1-antitrypsin disorder and for others awaiting transplant (as a multiple of the median for the group without liver disease) shows the differences in their anaemic profiles (Fig. 3).

Discussion

Cirrhosis of the liver in people with diabetes awaiting a liver transplant renders HbA_{1c} unsuitable for assessing blood glucose. In all but one person, it was associated with fewer, larger, more irregular red blood cells. A substantial depression in HbA_{1c} [20 mmol/mol (2%)] was observed relative to those with diabetes but no cirrhosis across a wide range of glucose values. This probably reflects a shorter red blood cell half-life and less exposure of haemoglobin to glucose. In contrast, the person with cirrhosis related to α-1-antitrypsin disorder had a higher HbA_{1c} level relative to glucose, with no factors indicating anaemia, suggesting the red blood cell half-life might be longer with more exposure to glucose.

This effect on HbA_{1c} in people with cirrhotic liver disease will cause misdiagnosis of diabetes and inappropriate clinical care. In our routine clinical practice, many more depressed

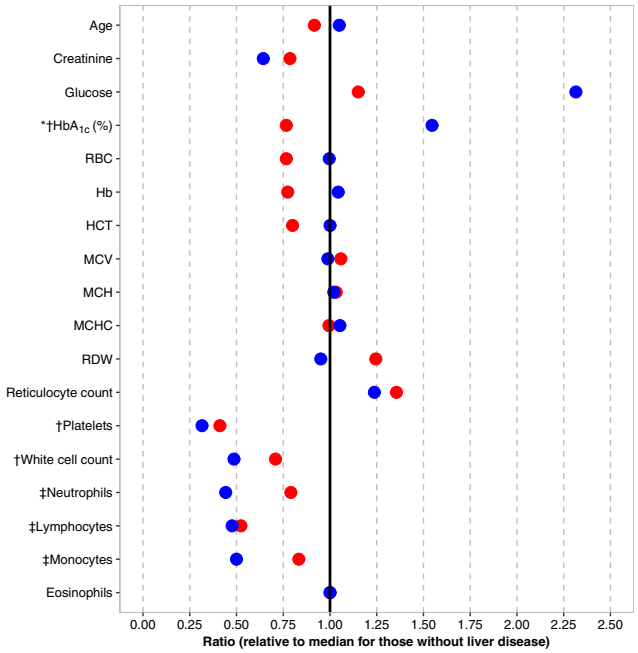


FIGURE 3 Comparison of haematology in people with diabetes and cirrhosis vs those with diabetes without liver disease. Circles: red for people with diabetes and cirrhosis, except for one person with α-1-antitrypsin disorder, which is blue. *P<0.01 for blue vs red; †P<0.01and ‡P<0.05 for blue vs people with diabetes without liver disease. RBC, red blood cell count; Hb, haemoglobin; HCT, haematocrit; MCV, mean cell volume; MCH, mean cell haemoglobin; MCHC, mean cell haemoglobin concentration; RDW, red cell distribution width.

than elevated HbA$_{1c}$ results have been noticed. We previously reported overtreatment resulting in hospital admission in one individual with known thalassaemia as a result of elevated HbA$_{1c}$ relative to glucose levels [12]. HbA$_{1c}$ assays do not identify thalassaemia, although some HbA$_{1c}$ analysers identify variant haemoglobins (e.g. S, F, C, D, E or rarer types) on chromatograms.

A recent US study in 200 people (62 with diabetes) referred for liver transplantation with decompensated cirrhosis showed similar depression in HbA$_{1c}$ relative to glucose [13]. HbA$_{1c}$ calculated from previous glucose results and compared to measured HbA$_{1c}$ [19], was found to be discordant by >0.5% in 49% of participants and >1.5% in 12% overall. Multivariate model analysis found haemoglobin to be the only independent predictor of the larger HbA$_{1c}$ discrepancies. More evidence is required regarding the extent of the effects of liver disease on the accuracy of HbA$_{1c}$ for clinical guidelines to improve on the diagnosis of diabetes and its management.

The groups differed distinctly when their biochemistry and haematology were compared, (Table 1 and Figs 2 and 3). The relationships of glucose and HbA$_{1c}$ to red blood cell haematology in people with diabetes and cirrhosis were markedly different from those in people with diabetes but no cirrhosis (Fig. 2). Low haemoglobin and macrocytosis evident in those with cirrhosis and diabetes were associated with depression in HbA$_{1c}$. The exception being the person with α-1-antitrypsin disorder whose erythrocytes did not display these features and whose HbA$_{1c}$ was elevated relative to glucose level. Anaemia can result in either shorter or longer erythrocyte lifespans and even differences in normal red blood cell morphology have been shown to affect the accuracy of HbA$_{1c}$ [20].

Any suspected inaccuracy in HbA$_{1c}$ can be confirmed using fructosamine, unless proteinuria is present [21], and point-of-care blood glucose testing or non-invasive continuous blood glucose devices. The data presented on >100 people attending the diabetes centre (along with corresponding fructosamine results) are used in our hospital to identify any outliers in glycaemic markers. As such, an elevated HbA$_{1c}$ relative to glucose level shows when additional testing, such as fructosamine/continuous blood glucose monitoring, should be organized by clinicians to confirm whether HbA$_{1c}$ is suitable for assessing glycaemic status. Monitoring glycaemia during the post-liver-transplant period is also an issue, as it is well known that post-transplant anaemia renders HbA$_{1c}$ unsuitable for clinical interpretation for ~6 months [5,22]. It is not known if this problem is resolved after liver transplantation.

Limitations of this study include the small number of people (29) studied with cirrhosis and diabetes compared to the available sample with diabetes but no cirrhosis (125). This sample size may hinder its ability to demonstrate statistical differences between the slopes of the regression lines. HbA$_{1c}$ was depressed by 25 mmol/mol (2.3%), (P<0.001), when the study was limited to age-matched white

European people with liver disease (mean age 55.6 years) compared to those without liver disease (mean age 55.3 years). As most of the participants were white European, it cannot throw any light on the current discussion about the relationship of HbA_{1c} to glucose by ethnicity [23]. Although random plasma glucose was measured rather than fasting, this reflects routine hospital practice as is evident in other studies [10]. Its measurement on glucose meters or blood gas machines quality-assured by the laboratory, or measured in the laboratory, is a quality indicator at the hospital. The number of people studied pre-transplant was small but it should be noted that the clinical audit was generated by observations of inaccurate HbA_{1c} in people with liver disease by experts in glycaemic markers over several years of routine clinical practice. Meta-analyses of small studies are common, with confirmatory studies required for clinical guidelines. Future research by our group will include more people with conditions that affect HbA_{1c} as outlined by WHO on more than one clinic visit [6].

In conclusion, cirrhosis of the liver affects the accuracy of HbA_{1c} results, leading to unreliable estimates of blood glucose over the previous 2 to 3 months. Anaemia in people with cirrhosis awaiting liver transplant is associated with altered red blood cell morphology. Significantly depressed HbA_{1c} was observed in all but one person with cirrhosis, along with lower haemoglobin level and fewer, larger, less uniform red blood cells. Visual representation of HbA_{1c} and random plasma glucose, along with haematology, is useful for assessing whether HbA_{1c} is accurate in individuals with coexisting illnesses or on drug regimens that affect red blood cells. Treatment targets for HbA_{1c} arising from clinical trials in diabetes [24,25] and cut-off values for diagnosis [7,23,26] rely on the provision of HbA_{1c} values that reflect circulating glucose.

Funding sources

The G. A. Roberts Research Fund, Queen Elizabeth Hospital Birmingham Charity, Tosoh Europe and Novo Nordisk UK Research Foundation provided funding for the Diabetes Translational Research Group based at Queen Elizabeth Hospital Birmingham. The Arthur Thompson Trust at the University of Birmingham Medical School and the Queen Elizabeth Hospital Birmingham Charity provided financial support for conference costs for D.B. and S.V. when these data were presented at the American Diabetes Association 74th Scientific Sessions, 2014 and Diabetes UK, 2015. Research by J.A.W. reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number UM1HG006370. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. G.V.G. acknowledges support from H2020-EINFRA (731075) and the National Science Foundation (IOS:1340112) as well as support from the NIHR Birmingham ECMC, NIHR Birmingham SRMRC and the NIHR Birmingham Biomedical Research Centre and the MRC HDR UK. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the Medical Research Council or the Department of Health. The funding organizations had no role in the design of this study, data collection, analysis or interpretation, or preparation of the manuscript, and did not approve or disapprove of, or delay publication of the work.

Competing interests

None declared.

Acknowledgements

The laboratory measurements were produced by the biomedical scientists in Clinical Laboratory Sciences at Queen Elizabeth Hospital Birmingham. We would like to thank Dr Radhika Susarla (Research Scientist, Diabetes Translational Research Group) for her assistance with the manuscript. We would also like to thank Dr Paul Cockwell (Consultant Nephrologist, Renal Medicine, Queen Elizabeth Hospital Birmingham) and Professor Wasim Hanif (Professor of Diabetes and Endocrinology, Diabetes Centre, Queen Elizabeth Hospital Birmingham) for their support and encouragement.

Data access statement

The datasets generated during and/or analysed during the study are not publicly available. The dataset contains clinical data which cannot be shared publicly as a result of UK data protection legislation.

References

1 Tolman KG, Fonseca V, Dalpiaz A, Tan MH. Spectrum of liver disease in type 2 diabetes and management of patients with diabetes and liver disease. *Diabetes Care* 2007; **30**: 734–743.
2 Ahmadieh H, Azar ST. Liver disease and diabetes: association, pathophysiology, and management. *Diabetes Res Clin Pract* 2014; **104**: 53–62.
3 Lecube A, Hernandez C, Genesca J, Simo R. Proinflammatory cytokines, insulin resistance, and insulin secretion in chronic hepatitis C patients: A case-control study. *Diabetes Care* 2006; **29**: 1096–1101.
4 Garcia-Compean D, Jaquez-Quintana JO, Gonzalez-Gonzalez JA, Maldonado-Garza H. Liver cirrhosis and diabetes: risk factors, pathophysiology, clinical implications and management. *World J Gastroenterol* 2009; **15**: 280–288.
5 Shivaswamy V, Boerner B, Larsen J. Post-Transplant Diabetes Mellitus: Causes, Treatment, and Impact on Outcomes. *Endocr Rev* 2016; **37**: 37–61.
6 Webber J, Chua S, Cockwell P, Haydon G, Jobanputra P, Lester W *et al.* Effects of concurrent illnesses and treatments on surrogate

glycaemic markers. *Diabet Med* 2017; **34** (Suppl. 1): P138 (Abstract).
7 Report of a World Health Organization Consultation. Use of glycated haemoglobin (HbA$_{1c}$) in the diagnosis of diabetes mellitus. *Diabetes Res Clin Pract* 2011;**93**:299–309.
8 Dowd RP, Manning PW, Ahmed N, Mason CL, Round RA, Nightingale PG *et al.* Post introduction of HbA1c as a diagnostic test: consequences for requesting and reporting. *Diabet Med* 2015; **32** (Suppl. 1): P440.
9 American Diabetes Association. Position statement. Diagnosis and classification of diabetes mellitus. *Diabetes Care* 2014; **37** (Suppl. 1): S81–S90.
10 Manley SE, O'Brien KT, Quinlan D, Round RA, Nightingale PG, Ali F *et al.* Can HbA$_{1c}$ detect undiagnosed diabetes in acute medical hospital admissions? *Diabetes Res Clin Pract* 2016; **115**: 106–114.
11 Dowd RP, Round RA, Mason CL, Nightingale PG, Ghosh SG, Hanif W *et al.* Review of HbA$_{1c}$ results >120 mmol/mol as patients may require urgent assessment if request for diagnosis of Type 2 diabetes. *Diabet Med* 2014; **31** (Suppl 1): P466.
12 Kadri F, Stuart K, Cramb R, Manley S, Mtemererwa B, Ghosh S. HbA$_{1c}$ interpretation and its caveats: a case of targeting apparently good glycaemic control causing severe hypoglycaemia. *International Diabetes Federation (IDF) meeting, Vancouver, Canada, 30 November to 4 December, 2015*. Abstract: 0550-P.
13 Nadelson J, Satapathy SK, Nair S. Glycated hemoglobin levels in patients with decompensated cirrhosis. *Int J Endocrinol* 2016; **2016**: Article ID 8390210. https://doi.org/10.1155/2016/8390210.
14 Malinchoc M, Kamath PS, Gordon FD, Peine CJ, Rank J, ter Borg PC. A model to predict poor survival in patients undergoing transjugular intrahepatic portosystemic shunts. *Hepatology* 2000; **31**: 864–871.
15 Manley SE, Hikin LJ, Round RA, Manning PW, Luzio SD, Dunseath GJ *et al.* Comparison of IFCC-calibrated HbA$_{1c}$ from laboratory and point of care testing systems. *Diabetes Res Clin Pract* 2014; **105**: 364–372.
16 R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2017. Available at https://www.R-project.org/. Last accessed 19 December 2017.
17 Wei T, Simko V. R package "corrplot": Visualization of a Correlation Matrix (Version 0.84) 2017. Available at https://github.com/taiyun/corrplot. Last accessed 19 December 2017
18 Revelle, W. psych: Procedures for Personality and Psychological Research (Version 1.7.8) 2017. Northwestern University, Evanston, Illinois, USA. Available at https://CRAN.R-project.org/package=psych. Last accessed 19 December 2017.
19 Nathan DM, Kuenen J, Borg R, Zheng H, Schoenfeld D, Heine RJ. Translating the A1C assay into estimated average glucose values. *Diabetes Care* 2008; **31**: 1473–1478.
20 Cohen RM, Franco RS, Khera PK, Smith EP, Lindsell CJ, Ciraolo PJ *et al.* Red cell life span heterogeneity in hematologically normal people is sufficient to alter HbA$_{1c}$. *Blood* 2008; **112**: 4284–4291.
21 Manley SE, Round RA, Nightingale PG, Stratton IM, Cramb R, Gough SC. How is fructosamine affected by urinary albumin excretion? *Diabetes* 2011; **60**(Suppl 1): A584.
22 Eide IA, Halden TA, Hartmann A, Åsberg A, Dahle DO, Reisæter AV *et al.* Limitations of hemoglobin A1c for the diagnosis of posttransplant diabetes mellitus. *Transplantation* 2015; **99**: 629–635.
23 American Diabetes Association. 2. Classification and diagnosis of diabetes: standards of medical care in diabetes-2018. *Diabetes Care* 2018; **41** (Suppl. 1): S13–S27.
24 DCCT Study Group. The relationship of glycemic exposure (HbA$_{1c}$) to the risk of development and progression of retinopathy in the diabetes control and complications trial. *Diabetes* 1995; **44**: 968–983.
25 UK Prospective Diabetes Study Group. UK Prospective Diabetes Study (UKPDS) 35. Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes : prospective observational study. *BMJ* 2000; **321**: 405–412.
26 Manley S, Nightingale P, Stratton I, Sikaris K, Smith J, Cramb R *et al.* Diagnosis of diabetes: HbA$_{1c}$ versus WHO criteria. *Diabetes Prim Care* 2010; **12**: 87–96.

Received: 11 March 2020 | Accepted: 4 April 2020

DOI: 10.1002/edm2.140

ORIGINAL RESEARCH ARTICLE

Endocrinology, Diabetes & Metabolism Open Access WILEY

Prevalence of admission plasma glucose in 'diabetes' or 'at risk' ranges in hospital emergencies with no prior diagnosis of diabetes by gender, age and ethnicity

Sandip Ghosh[1,2] | Susan E. Manley[1,2,3] | Peter G. Nightingale[2] | John A. Williams[2,4,5] | Radhika Susarla[1,2] | Irene Alonso-Perez[6] | Irene M. Stratton[7] | Georgios V. Gkoutos[2,4,8,9] | Jonathan Webber[1] | Stephen D. Luzio[1,10] | Wasim Hanif[1,2] | Graham A. Roberts[1,10,11]

[1]Diabetes Translational Research Group, Diabetes Centre, Nuffield House, Queen Elizabeth Hospital Birmingham, Birmingham, UK

[2]Institute of Translational Medicine, Heritage Building (Queen Elizabeth Hospital), Birmingham, UK

[3]Institute of Metabolism and Systems Research, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK

[4]Institute of Cancer and Genomic Sciences, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK

[5]Mammalian Genetics Unit, Medical Research Council Harwell Institute, Oxfordshire, UK

[6]Health Informatics Department, Queen Elizabeth Hospital Birmingham, Birmingham, UK

[7]Gloucestershire Retinal Research Group, Cheltenham General Hospital, Cheltenham, UK

[8]MRC Health Data Research UK (Central Office), Gibbs Building, London, UK

[9]NIHR Biomedical Research Centre, Birmingham, UK

[10]Diabetes Research Unit (Cymru), Grove Building, Swansea University, Swansea, UK

[11]HRB-Clinical Research Facility – Cork, Mercy University Hospital, Cork, Ireland

Correspondence
Susan E. Manley, Diabetes Centre,

Abstract

Aims: To establish the prevalence of admission plasma glucose in 'diabetes' and 'at risk' ranges in emergency hospital admissions with no prior diagnosis of diabetes; characteristics of people with hyperglycaemia; and factors influencing glucose measurement.

Methods: Electronic patient records for 113 097 hospital admissions over 1 year from 2014 to 2015 included 43 201 emergencies with glucose available for 31 927 (74%) admissions, comprising 22 045 people. Data are presented for 18 965 people with no prior diagnosis of diabetes and glucose available on first attendance.

Results: Three quarters (14 214) were White Europeans aged 62 (43-78) years, median (IQ range); 12% (2241) South Asians 46 (32-64) years; 9% (1726) Unknown/Other ethnicities 43 (29-61) years; and 4% (784) Afro-Caribbeans 49 (33-63) years, $P < .001$. Overall, 5% (1003) had glucose in the 'diabetes' range (≥11.1 mmol/L) higher at 8% (175) for South Asians; 16% (3042) were 'at risk' (7.8-11.0 mmol/L), that is 17% (2379) White Europeans, 15% (338) South Asians, 14% (236) Unknown/Others and 11% (89) Afro-Caribbeans, $P < .001$. The prevalence for South Asians aged <30 years was 2.1% and 5.2%, respectively, 2.6% and 8.6% for Afro-Caribbeans <30 years, and 2.0% and 8.4% for White Europeans <40 years. Glucose increased with age and was more often in the 'diabetes' range for South Asians than White Europeans with South Asian men particularly affected. One third of all emergency admissions were for <24 hours with 58% of these having glucose measured compared to 82% with duration >24 hours.

Conclusions: Hyperglycaemia was evident in 21% of adults admitted as an emergency; various aspects related to follow-up and initial testing, age and ethnicity need

Sandip Ghosh and Susan E. Manley are equal first authors.

Stephen D. Luzio, Wasim Hanif, and Graham A. Roberts are equal last authors.

Nuffield House, Queen Elizabeth Hospital Birmingham, Mindelsohn Way, Edgbaston, Birmingham B15 2GW, UK.
Email: susan.manley@uhb.nhs.uk

Funding information
This study was performed by the Diabetes Translational Research Group and supported by: G. A. Roberts Research Fund and Queen Elizabeth Hospital Birmingham Charity. JAW received grants from National Human Genome Research Institute of National Institutes of Health, under award number UM1HG006370. GVG received grants from H2020-EINFRA (731075), National Science Foundation (IOS:1340112), NIHR Birmingham ECMC, NIHR Birmingham SRMRC, NIHR Birmingham Biomedical Research Centre and MRC HDR UK. The views expressed in this paper are those of the authors and not necessarily those of the NHS, National Institute for Health Research, Medical Research Council, Department of Health or US National Institutes of Health. The funding organizations had no role in the design of the study, data collection, analysis or interpretation, or preparation of manuscript, and did not approve/disapprove of, or delay publication.

to be considered by professional bodies addressing undiagnosed diabetes in hospital admissions.

KEYWORDS
emergency admissions, hyperglycaemia, undiagnosed diabetes

1 | INTRODUCTION

The current diabetes pandemic threatens both the health and economy of nations.[1] The prevalence is increasing year by year adding markedly to the cost of health care funded by governments or private healthcare organizations. Diabetes currently consumes over 10% of the UK National Health Service (NHS) budget.[2] The prevalence of type 2 diabetes and its complications vary by ethnicity with type 2 diabetes more prevalent in people of South Asian descent (six times) and in Africans and Afro-Caribbeans (three times) than White Europeans.[3,4]

Symptoms of diabetes are not always evident until people consult a family doctor or are admitted to hospital.[5] Undiagnosed diabetes results in an eightfold increase in mortality for hospital admissions compared to those with normal glucose.[6] Admission glucose is strongly associated with mortality in acutely ill medical patients[7] with hyper/hypoglycaemia independent predictors of in-hospital mortality in patients not previously diagnosed with diabetes.[8,9]

The American Diabetes Association guidance published in 2020 recommends measuring HbA1c in patients admitted with hyperglycaemia, defined as glucose >7.8 mmol/L.[6] Although HbA1c is used for diagnosis in the community in the UK following WHO guidance in 2011,[10] it is not currently requested routinely on hospital admission for this purpose. Data from an Irish hospital indicates that HbA1c could be used for follow-up testing[11] although it is not suitable for people with some haemoglobinopathies or altered red blood cell turnover.[10]

This clinical audit reports on admission plasma glucose in emergency admissions with no prior diabetes coding using laboratory and demographic data from hospital electronic patient records over 1 year. It aims to investigate people with glucose recorded on admission and no prior diabetes diagnosis, to categorize plasma glucose by 'at risk' and 'diabetes' ranges, and describe relationships to age, gender and ethnicity.

2 | METHODS

2.1 | Design

The clinical audit at Queen Elizabeth Hospital Birmingham was approved according to clinical governance (*CARMS-12031*), University Hospitals Birmingham NHS Foundation Trust. This hospital, located in the West Midlands of England with a culturally diverse catchment, is a major trauma centre for adults comprising >1200 beds with 100 critical care beds.

2.2 | Admissions

There were 113 097 admissions between April 2014 and March 2015 with 38% (43 201) emergencies, others elective admissions or day care patients (Figure 1). Self-reported ethnicity was coded as White European/Caucasian (British, Irish or any other white background); South Asian (Indian, Pakistani, Bangladeshi or any other Asian background); Afro-Caribbean (Caribbean, African or any other black background); other groups (Chinese, any other ethnic groups not described above, and mixed ethnic groups)

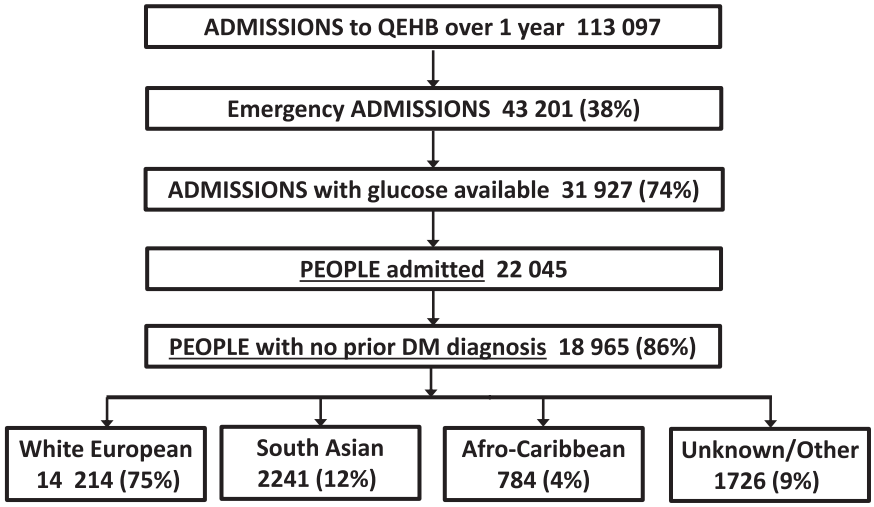


FIGURE 1 Flow chart for emergency admissions to a UK hospital located in a multi-ethnic region over 1 y: 2014-2015

and unknown (those not specified and missing). Diabetes status was assigned on admission from ICD 10 coding (International Classification of Diseases, Tenth Revision). People with multiple admissions were identified, and their first admission designated as the index admission.

2.3 | Study data

Data were obtained from the electronic patient record system PICS (patient information and communication system) with the initial plasma glucose result measured by point-of-care testing or in the laboratory.

2.4 | Measurements

Blood glucose was measured at point-of-care on the wards in random capillary blood on glucose meters or arterial/venous whole blood samples on gas machines with results reported as plasma; Roche Cobas Inform II glucose meters (CV 5%) and Roche Cobas b221 blood gas machines (CV 4%). Blood was collected into fluoride oxalate vacutainers for measurement of plasma glucose in the central hospital laboratory using Roche Cobas 8000 analysers (CV 2%). Internal quality control and external quality assurance were overseen by the hospital blood sciences laboratory. The performance of the point-of-care equipment was compared with the laboratory analysers and found to be acceptable.

2.5 | Statistical analysis

Anonymized clinical audit data were downloaded by Health Informatics, stored and analysed by a hospital statistician and data visualization analyst using Microsoft Excel, IBM SPSS Statistics for Windows version 22.0 (IBM Corp.) and R version 3.4.0 including ggplot2 and vcd packages.[12-14]

Variables are presented in Tables 1 and 2 as median and interquartile range, or count/percentage. Mann-Whitney or Fisher's exact tests were used to compare groups. The association between the prevalence of prior diabetes diagnosis coding and number of admissions was assessed using Kendall's tau-*b* statistic.

Age bands of 15-19 years, 20-24 years, 25-29 years up to 95-99 years and 100-104 years were used to construct Figure 3A which shows the proportion of patients in each age band by ethnicity. For Figure 3B,C, predicted glucose values obtained from a linear regression model of log glucose on age, sex and ethnicity (including interactions, Table 3) were plotted against age. Differences between the sexes and between ethnic groups were assessed by examining their interaction with age. No adjustment was made for multiple comparisons.

Bonferroni-corrected *t* tests were performed to investigate age and gender differences in admission plasma glucose in non-repeat admissions within each ethnic group. Multilevel contingency tables associating frequencies of glucose, sex, and ethnicity (Figure 4A), and glucose, sex, and age group within the ethnic groups (Figure 4B-E) were analysed with Pearson's chi-square test for independence.

Power calculations were performed using sample sizes for people in ethnic groups in Table 2 with a significance level of .05/6 used to adjust for multiple comparisons. Due to unequal sample size, the power to detect a difference depended on whether the lower proportion related to the smaller or larger group.

3 | RESULTS

3.1 | People admitted as an emergency

Of the 113 097 admissions over 1 year, 38% (43 201) were emergency (Figure 1). Plasma glucose was measured on admission in 74% (31 927) (Table 1), with 75% from blood glucose meters, 18% blood gas machines and 7% laboratory analysers. Out of the 5867 admissions with a prior diabetes diagnosis, 94% (5523) had plasma glucose reported.

TABLE 1 Admission plasma glucose in emergency admissions over 1 y from 2014 to 2015

	All emergency admissions[a]	Admissions with glucose available[b]	People with glucose available	People with glucose available and no prior diabetes coding[b,c]
n	43 201	31 927 (74%)	22 045	18 965 (86%)
Age[d], y	60 (41-77)	62 (43-77)[e,***]	60 (41-76)	58 (38-76)[e,***]
Female	21 658 (50%)	15 947 (74%)	10 936	9539 (87%)***
Male	21 233 (50%)	15 744 (74%)	10 982	9325 (85%)
White European (WE)	32 581 (75%)	23 942 (73%)	16 271	14 214 (87%)
South Asian (SA)	5408 (13%)	4243 (78%)***	2910	2241 (77%)***
Unknown/Other (U/O)	3420 (8%)	2351 (69%)***	1891	1726 (91%)***
Afro-Caribbean (AC)	1792 (4%)	1391 (78%)***	973	784 (81%)***
Diabetes coding				
Prior	5867 (14%)	5523 (94%)***	3080	—
No prior	37 334 (86%)	26 404 (71%)	18 965	18 965 (100%)
Admission <24 h	14 181 (33%)	8258 (58%)***	6224	5524 (89%)***
Admission ≥24 h	29 020 (67%)	23 669 (82%)	15 821	13 441 (85%)
Repeat admission	12 537 (29%)	9882 (79%)***	—	—
Glucose[d] mmol/L	—	6.4 (5.4-8.0)	6.4 (5.4-7.9)	6.2 (5.3-7.4)***

[a]% for categories within column.

[b]*P* values for comparing % in each category; WE, reference category for ethnicity.

[c]*P* values for comparison of glucose for column 4 vs those in 3 but not 4.

[d]Median and quartiles otherwise n (%).

[e]*P* values for comparison of age—column 2 vs those in 1 but not 2 and column 4 vs those in 3 but not 4.

***$P < .001$.

TABLE 2 Ethnic differences in people admitted as an emergency with glucose measured on admission but no prior diagnosis of diabetes

	People with no prior diabetes coding and glucose available[a]	White European (WE)	South Asian (SA)	Unknown/ Other (U/O)	Afro-Caribbean (AC)	*P* values
n	18 965	14 214	2241	1726	784	
Age[b], y	58 (38-76)	62 (43-78)	46 (32-64)	43 (29-61)	49 (33-63)	c***
Age ≥90 y	968 (5%)	906 (6%)	31 (1%)	24 (1%)	7 (1%)	
Female, n (%)	9539 (50%)	7245 (51%)	1136 (51%)	761 (44%)	397 (51%)	d***
Male	9325 (49%)	6969 (49%)	1105 (49%)	864 (50%)	387 (49%)	
Not recorded	101 (1%)	0 (0%)	0 (0%)	101 (6%)	0 (0%)	
Glucose[b] mmol/L	6.2 (5.3-7.4)	6.2 (5.4-7.5)	6.2 (5.3-7.5)	5.9 (5.2-7.2)	6.0 (5.2-7.2)	e***
Ranges, n (%)						
<5.0	2672 (14%)	1868 (13%)	364 (16%)	291 (17%)	149 (19%)	
5.0-5.5	3157 (17%)	2309 (16%)	375 (17%)	332 (19%)	141 (18%)	
5.6-7.7	9091 (48%)	6970 (49%)	989 (44%)	778 (45%)	354 (45%)	
7.8-11.0	3042 (16%)	2379 (17%)	338 (15%)	236 (14%)	89 (11%)	
>11.0	1003 (5%)	688 (5%)	175 (8%)	89 (5%)	51 (7%)	f***

[a]Single/index if multiple admissions.

[b]Median, IQ range.

***$P < .001$ for [c]WE vs SA, WE vs U/O, WE vs AC, SA vs U/O, U/O vs AC; [d]WE vs U/O after excluding 'not recorded'; [e]WE vs U/O & AC, SA vs U/O & AC (WE & SA not significantly different nor U/O & AC); [f]for proportion with glucose >11.0 mmol/L for WE vs SA, U/O vs SA.

3.2 | Availability of glucose

One third of all emergency admissions stayed in hospital for <24 hours and two thirds for >24 hours; 58% of emergency admissions with a duration <24 hours had glucose available and 82% of emergency admissions with duration >24 hours (Table 1). Admissions with plasma glucose reported (31 927) were 2 years older than the total cohort (43 201), median, IQ range for age, 62 (43-77) vs 60 (41-77) years, $P < .001$ (Table 1),

TABLE 3 Equations relating glucose in mmol/L to age in years for each sex for the different ethnic groups

	White European (WE)	South Asian (SA)	Unknown/Other (U/O)	Afro-Caribbean (AC)
Males, n	6966	1104	863	387
Females, n	7244	1136	761	397
Males equation	$\log_{10}$ glucose = 0.00094 × age + 0.761	$\log_{10}$ glucose = 0.00177 × age + 0.743	$\log_{10}$ glucose = 0.00143 × age + 0.743	$\log_{10}$ glucose = 0.00105 × age + 0.758
P value	<.001[a]	<.001[b]	<.05[b]	<.05[a]
Females equation	$\log_{10}$ glucose = 0.00129 × age + 0.732	$\log_{10}$ glucose = 0.00230 × age + 0.694	$\log_{10}$ glucose = 0.00147 × age + 0.720	$\log_{10}$ glucose = 0.00229 × age + 0.686
P value	—	<.001[b]	—	<.01[b]

[a]Comparison of age coefficient to that for females from the same ethnic group.
[b]Comparison of age coefficient to that for White Europeans of the same sex.

with no differences in gender. Admissions without glucose measurements were 8 years younger at 54 (36-73) years, *P* < .001.

A substantial number of emergency admissions, 12 537, were re-admissions and glucose was measured in 79% (9882) of these. A total of 30 664 people were admitted with 76% (23 411) admitted once and 24% (7253) readmitted. Glucose was not measured on admission in 28% (8619) of these people who were younger at 53 (34-71) years, than those with glucose available, 72% (22 045), aged 60 (41-76) years, *P* < .001.

3.3 | Timing of admission

Glucose was more likely to be measured on admission to Queen Elizabeth Hospital Birmingham at the weekend (Saturday/Sunday), 76% vs 73%, *P* < .001, or during the week between 6 PM and 6 AM, 76% vs 71%, *P* < .001.

3.4 | Multiple admissions

People readmitted (7253) were 9 years older than those admitted once (23 411), *P* < .001 with a higher proportion of women, 52% vs 49%, and White Europeans, 79% vs 73%, and fewer from Unknown/Other ethnic groups, 5% vs 10%, all at *P* < .001. Of these re-admissions, 63% (4569) were on two occasions, 20% (1472) three, 9% (630) four and 8% (582) on five or more occasions. People readmitted were more likely to be coded for diabetes, 17% (1266) vs 9% (2045), *P* < .001. The prevalence of prior diabetes increased with the number of admissions, that is 15% (689) for two, 18% (272) for three, 23% (146) for four and 27% (159) for five or more admissions, Kendall's tau-*b* = 0.09, *P* < .001.

3.5 | Glucose and glycaemic status on admission

Glucose was 6.4 (5.4-8.0) mmol/L in 74% (31 927) of admissions with glucose available (Table 1); 8.8 (6.6-12.5) mmol/L in 17% (5523) of these admissions with prior diabetes coding and 6.2 (5.3-7.4) mmol/L in 83% (26 404) with no diabetes coding. In admissions without prior diabetes coding, 31% (8059) were ≤5.5 mmol/L, 48% (12 704) 5.6-7.7 mmol/L, 16% (4283) 7.8-11.0 mmol/L, that is 'at risk' range and 5% (1358) ≥11.1 mmol/L 'diabetes' range.

Over 20% of the people admitted as an emergency had hyperglycaemia; 5% had glucose in the 'diabetes' range and 16% in the 'at risk' range, Table 2 and Figure 2 with a higher proportion of South Asians (8%) than White Europeans (5%) in the 'diabetes' range, *P* < .001. The proportion of White Europeans (17%) and South Asians (15%) in the 'at risk' range was higher than for Afro-Caribbeans, (11%) *P* < .001 and *P* = .010. Some guidance specifies age limits below which people should not be tested for undiagnosed diabetes.[15] For South Asians aged <30 years, glucose was in the 'diabetes' and 'at risk' ranges for 2.1% and 5.2%, respectively, for Afro-Caribbeans aged <30 years 2.6% and 8.6%, for Unknown/Others aged <40 years 1.9% and 9.0%,

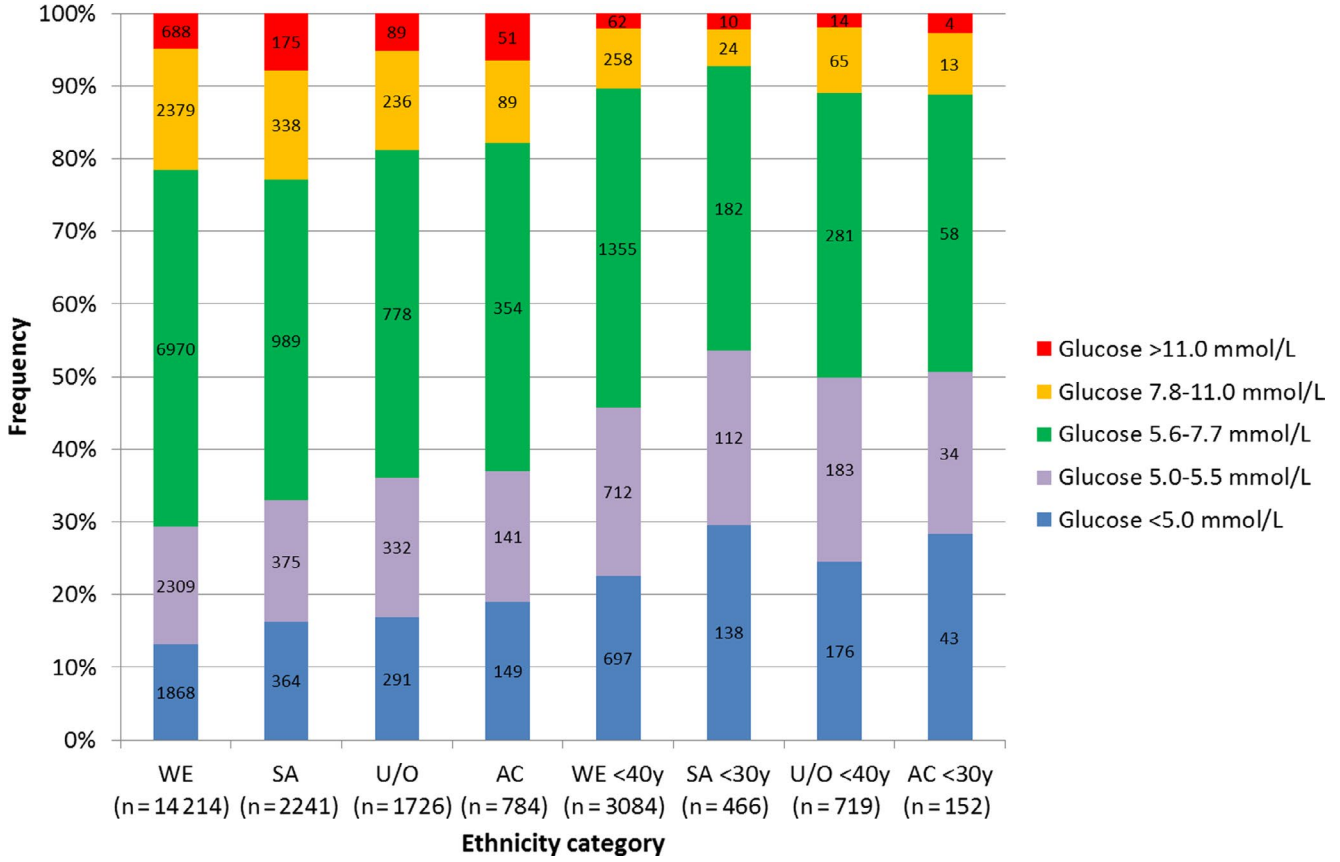


FIGURE 2 Glucose ranges for people admitted as an emergency without prior diabetes diagnosis by ethnicity and proposed age limit for follow-up. AC, Afro-Caribbean; SA, South Asian; U/O, Unknown/Other; WE, White European

and for White Europeans aged <40 years 2.0% and 8.4%. When people below these age limits were excluded, the prevalences were 6.3% and 18.4% compared with 5.3% and 16.0% when all ages were included.

3.6 | Ethnicity of people admitted

Three quarters of people admitted without prior diabetes coding with glucose available were White European, median age 62 years (Table 2). They were older than 12% of South Asians (46 years), 9% when ethnicity Unknown/Other, (43 years) and 4% Afro-Caribbean (49 years), $P < .001$. However, the age distributions of the ethnic groups were markedly different (Figure 3A). Overall, 5% (968) of people were aged 90 years old or older with 6% White European and 1% from the other ethnic groups (Table 2). Proportionally glucose was available for more South Asian and Afro-Caribbean admissions, 78% vs 73% for White European and 69% for Unknown/Others (Table 1).

3.7 | Admission glucose by age, gender and ethnicity

Glucose was higher as the age of the people admitted increased (Figure 3B,C), and varied by ethnicity. South Asian men aged >21 years and women aged >37 years had higher glucose than the White Europeans. In terms of overall plasma glucose levels, White Europeans and South Asians had slightly higher median glucose on admission at 6.2 mmol/L than Afro-Caribbeans, 6.0 mmol/L, and Unknown/Others 5.9 mmmol/L, $P < .001$. There were significant differences in glucose over the age distribution depending on people's gender and ethnicity (Table 3). Increases in glucose with age were greater for South Asian men and women than for White Europeans of the same gender; also for men whose ethnicity was Unknown/Other and Afro-Caribbean women (Table 3).

When relating ethnicity to glucose ranges, South Asian men were more prevalent than expected in the 'diabetes' range, >11.0 mmol/L, and South Asian women in the lowest range, <5.0 mmol/L (Figure 4A). A similar analysis with age showed that South Asian women aged <30 years were most prevalent in the lowest glucose range, with South Asian men and women aged >70 years more prevalent in the 'diabetes' and 'at risk' ranges (Figure 4C). White European men aged 50-69 years old were prevalent in the 'diabetes', 'at risk' and 'prediabetes' ranges (Figure 4B).

3.8 | Diabetes diagnosed during hospital admission

Diabetes was diagnosed following routine protocol during the index admission in people without a prior diagnosis before

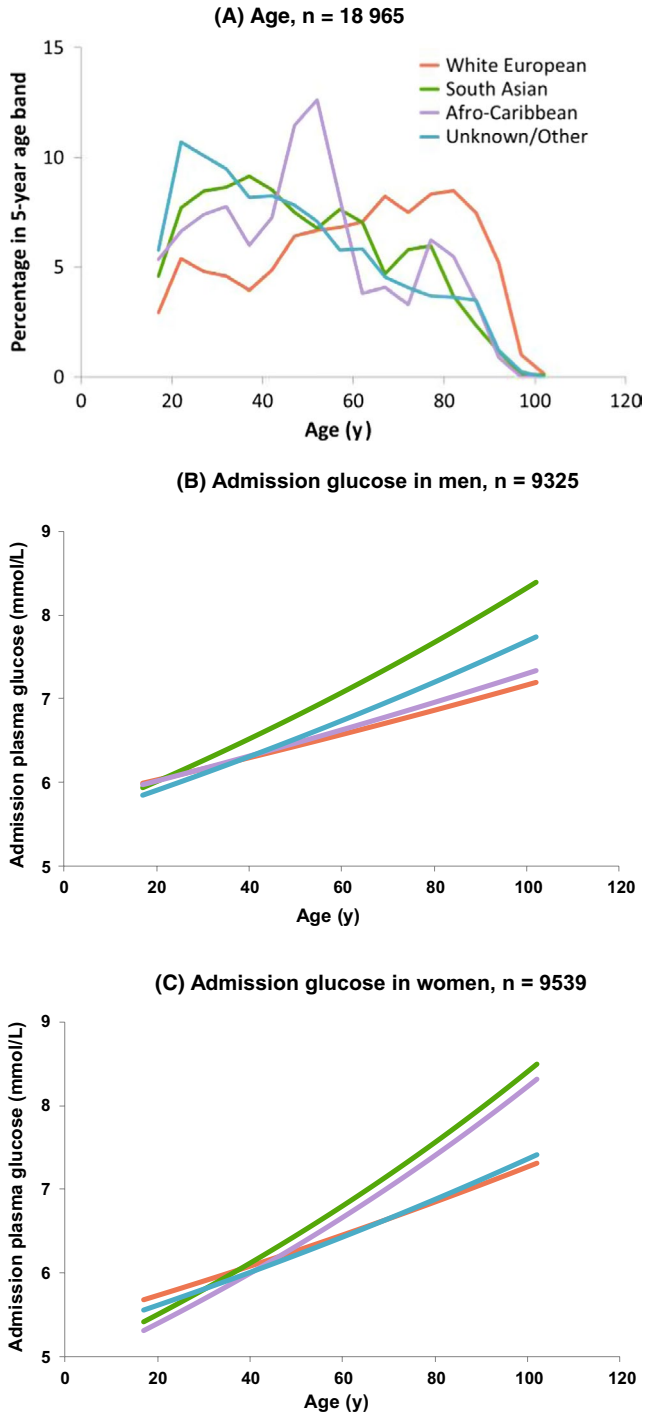


FIGURE 3 Age distribution and predicted admission glucose of people admitted to hospital as an emergency without prior diabetes diagnosis by ethnicity. Purple: Afro-Caribbean; green: South Asian; blue: Unknown/Other ethnic groups; orange: White European

admission (and with glucose measured) in 10% (1849/18 965), that is in 8% (1184/14 214) of White Europeans, 10% (168/1726) Unknown/Other ethnic groups, 14% (108/784) Afro-Caribbeans and 17% (389/2241) South Asians. Of the 1849 newly diagnosed, 575 were in the 'diabetes' range (31%) and 495 were in the 'at risk' range (27%).

4 | DISCUSSION

The prevalence of diabetes recorded in people in hospital in Birmingham is nearly double that of the community at 22% vs 12%.[16] Early diagnosis of diabetes is important as people can be advised to alter their diet, exercise regimen and lifestyle, or blood glucose lowering treatment can be introduced when necessary.

Hyperglycaemia on admission to hospital is defined as 'at risk' of diabetes, that is 7.9–11 mmol/L or in the 'diabetes' range, that is >11 mmol/L. Immediate action is required on the ward when people present with very high glucose, for example 25/30 mmol/L[15] to prevent/diagnose life-threatening conditions such as diabetic ketoacidosis.

Undiagnosed diabetes may be the cause of hyperglycaemia on admission but its diagnosis should be confirmed by additional testing with HbA1c. In this audit, 5% of White Europeans and 8% of South Asians and Afro-Caribbeans had glucose in the 'diabetes' range but there is little evidence on how many cases of diabetes would be confirmed on HbA1c testing. People in this study below the age limits specified in some guidance on additional testing had glucose in the abnormal ranges.

National protocols for identifying undiagnosed diabetes in admissions are mainly based on expert opinion and do not address the entire process from flexi-testing in a hospital laboratory to follow-up by GPs. Medico-legal implications can arise when abnormal glucose is not acted on during admission as people may present some years later with diabetes complications if not diagnosed during or after their hospital stay (Personal communication from Dr Sandip Ghosh and Professor Graham Roberts). Some preliminary data on those diagnosed with diabetes on admission using routine procedures are presented here. But, it requires more attention by the research team as it could reflect coding practice and is included in an ongoing project.

On emergency admission, 94% of those admitted to this hospital with a previous diabetes diagnosis had glucose recorded, but the figure for those not previously diagnosed was 74%. How this performance compares with other UK hospitals could be assessed by national inpatient audit programs. This figure may be related to the length of stay in hospital. One third of all emergency admissions were for <24 hours with 58% of these having glucose measured compared to 82% with a duration of >24 hours (Table 1). Those without glucose available were younger and more likely to be White European. The time/day of admission did not markedly influence the availability with only small differences observed possibly reflecting the hospital organization.

The number of people in the 'at risk' range was much higher than those in the 'diabetes' range involving 16% of people admitted with glucose measured. However, this audit is limited by the length of time the various ethnic groups have resided in the West Midlands.[17] As South Asians and Afro-Caribbeans present with diabetes at a younger age than White Europeans, it is vital to consider the age cut-offs for further testing quoted in some cases as 30 years for the former groups and 40 years for latter.[15] In routine practice, we have identified South Asian males presenting in their 20s with very high glucose and diabetic ketoacidosis—the reason why this audit was generated. As there were fewer Afro-Caribbeans in the audit, the power to

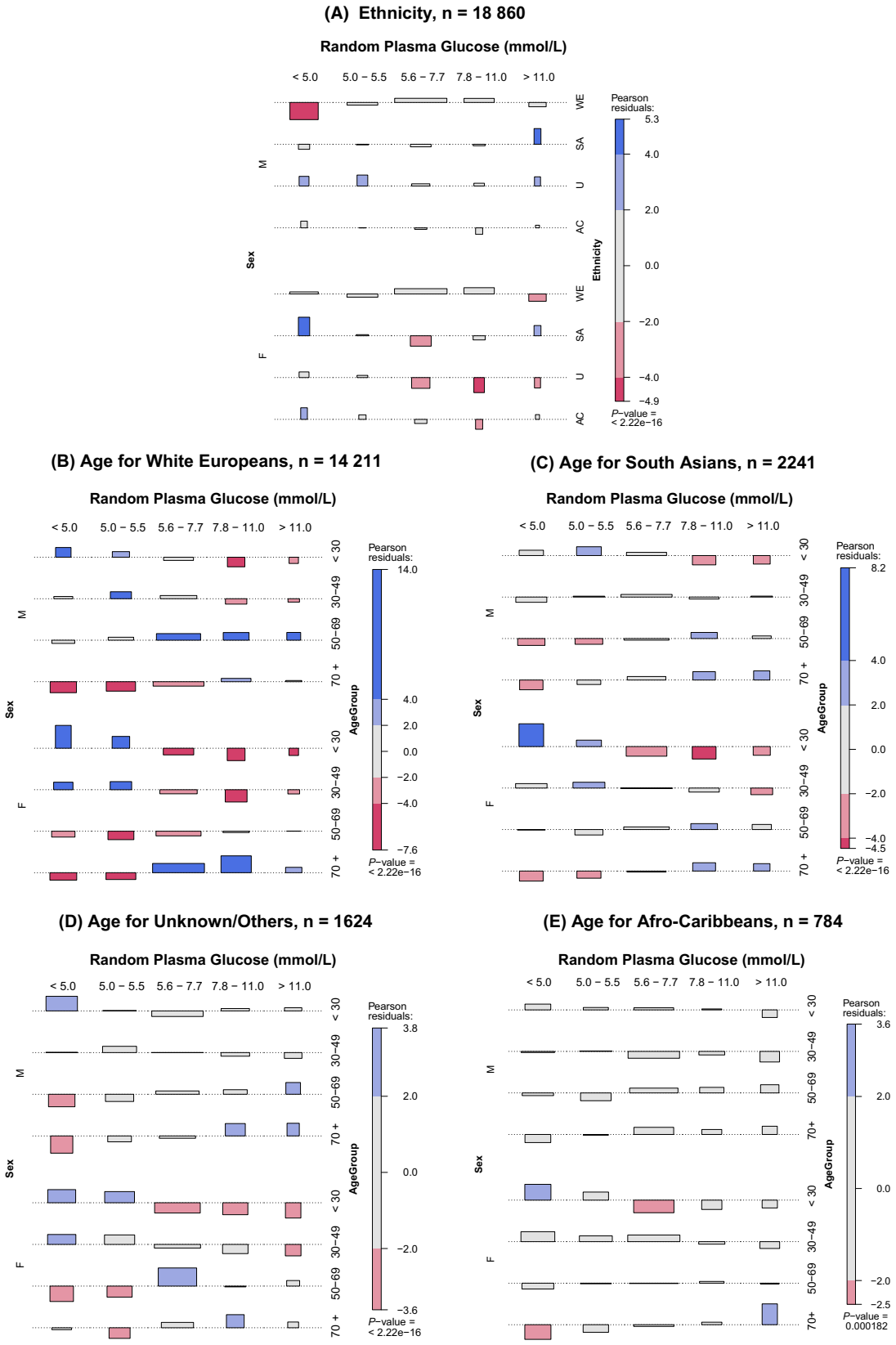


FIGURE 4 Glucose levels in people admitted to hospital as an emergency without prior diabetes diagnosis by gender, ethnicity and age. M, male; F, female. Blue—frequency observed significantly more than expected with positivity increasing with depth of colour; red—less frequently; grey—nonsignificant residuals. Boxes—areas proportional to difference in observed and expected frequencies; dashed grey baseline is expected count; above baseline greater than expected frequencies; below fewer than expected frequencies. Height is proportional to contribution of Pearson's residuals; width is proportional to square root of expected counts

detect differences between proportions of 5% and 8% in this group was lower at 54%-78% than for the other ethnic groups, 87%-100%.

The data presented here suggest that there should be no lower limit for follow-up testing in adults (Figure 2). This is important as diagnosis impacts on both disease progression and the risk of developing complications which are apparent several years before diagnosis. This evidence will help guideline writers to assess the workload and cost of implementing procedures for diagnosing diabetes in emergency admissions. The Joint British Diabetes Societies (JBDS) guidance on diabetes at the front door issued in February 2020[15] recommends further testing with HbA1c if glucose >7.8 mmol/L in people aged 40 years or older or 30 years depending on ethnicity.

Upgrading electronic patient record systems to identify people with glucose in 'diabetes' range and to prioritize further HbA1c testing could help reduce readmission rates. As raised glucose on admission was a particular problem in re-admissions, only their first admission was included in subsequent analyses. The prevalence of admission glucose in the 'diabetes' range was 27% for those admitted on five or more occasions.

Medical history/current records should be accessed when following up raised admission glucose as conditions requiring emergency hospitalization can cause anaemia which may affect the accuracy of HbA1c.[18,19] Ethnic differences in its relationship with glucose may be linked to red blood cell morphology.[20,21] In addition, variation in people with normal haematological profiles can account for differences of up to 5 mmol/mol (0.5%). Some countries are questioning whether different HbA1c cut-offs are necessary for diagnosis of diabetes. When inaccuracy is suspected, fructosamine can confirm abnormal glucose levels but the test is not recommended for diagnosis.

HbA1c is requested on admission of people without diagnosed diabetes now in some hospitals in Europe, America and Australia but published data on its efficacy is minimal.[6,22] When diagnosis using HbA1c was compared with OGTT if fasting glucose raised in general practice, correlation on diabetes diagnosis reached 95% when HbA1c >57 mmol/mol (7.5%).[23] A recent study of Australian adults aged ≥60 years reports a low diagnosis rate for diabetes in emergency hospital admissions due to people going into hospital undiagnosed and remaining undiagnosed during admission or with HbA1c results not necessarily communicated to family doctors on discharge.[24]

5 | CONCLUSIONS

A significant number of people admitted as an emergency but not previously diagnosed with diabetes had hyperglycaemia within the 'diabetes' (5%) and 'at-risk' ranges (16%) (Figure 2). South Asians were admitted at a younger age than White Caucasians with their admission glucose higher and South Asian men particularly affected (Figure 4A). This audit highlighted various issues regarding the availability of glucose on admission (75%), readmission rate as hyperglycaemia increased with the number of admissions, whether age limits should be employed for additional HbA1c testing to confirm diagnosis as people below limits specified had admission glucose in the abnormal ranges, and the effect of the length of time the various ethnic groups have resided in the West Midlands. Further investigation into the efficacy, procedures and cost of diagnosis in emergency admissions is required—this will involve reflex HbA1c testing and algorithms linking hospital and primary care. Liaison between public health, diabetes organizations and researchers is required to address these issues.

ACKNOWLEDGEMENTS

Laboratory measurements were produced by the biomedical scientists in Clinical Laboratory Sciences at Queen Elizabeth Hospital Birmingham. We would like to thank G. Gill (University Hospitals Birmingham NHS Foundation Trust) who performed the audit and R. A. Round (University Hospitals Birmingham NHS Foundation Trust) for their assistance with the manuscript.

CONFLICT OF INTEREST

There are no conflicts of interest for the authors. The study sponsor, University Hospitals Birmingham NHS Foundation Trust, was not involved in the design of the study; the collection, analysis and interpretation of data; writing the report; or the decision to submit the report for publication.

AUTHOR CONTRIBUTIONS

All authors qualify for authorship based on the International Committee of Medical Journal Editors criteria. All authors take full responsibility for content of manuscript. SG designed and organized the clinical audit and reviewed data and manuscript. SEM was responsible for data analysis and writing the manuscript. I.A-P. created the database. PGN and JAW analysed data and edited the manuscript. IMS and GVG reviewed the data analyses and the manuscript. RS contributed to data analysis and interpretation, and writing the manuscript. JW reviewed clinical aspects and also the manuscript. SDL contributed to data analysis and interpretation, and reviewed the manuscript. GAR and WH contributed to the overall audit process and reviewed clinical aspects of the paper. SG is the guarantor of this work.

DATA AVAILABILITY STATEMENT

The data sets generated during and/or analysed during the study are not publicly available. The data set contains clinical data which cannot be shared publicly due to UK data protection legislation.

ORCID

Sandip Ghosh https://orcid.org/0000-0003-0333-5992
Susan E. Manley https://orcid.org/0000-0002-8298-4511
John A. Williams https://orcid.org/0000-0002-0357-5454
Radhika Susarla https://orcid.org/0000-0002-0492-6519
Irene M. Stratton https://orcid.org/0000-0003-1172-7865
Georgios V. Gkoutos https://orcid.org/0000-0002-2061-091X
Stephen D. Luzio https://orcid.org/0000-0002-7206-6530
Graham A. Roberts https://orcid.org/0000-0002-5018-0391

REFERENCES

1. International Diabetes Federation. *IDF Diabetes Atlas*, 9th edn. Brussels; 2019. https://www.diabetesatlas.org. Accessed March 10, 2020.
2. Action for Diabetes – NHS England. https://www.england.nhs.uk/rightcare/wp-content/uploads/sites/40/2016/08/act-for-diabetes-31-01.pdf. Accessed March 10, 2020.
3. Black SA. Diabetes, diversity, and disparity: what do we do with the evidence? *Am J Public Health*. 2002;92:543-548.
4. Spanakis EK, Golden SH. Race/ethnic difference in diabetes and diabetic complications. *Curr Diab Rep*. 2013;13:814-823.
5. Rahman N, Collins A, Sheehan J, Perry I. Undetected hyperglycaemia among hospital in-patients. *Ir Med J*. 2000;93:268-270.
6. American Diabetes Association. 14. Diabetes Care in the Hospital: Standards of Medical Care in Diabetes-2018. *Diabetes Care*. 2020;41(Suppl 1):S144-S151.
7. Akirov A, Diker-Cohen T, Masri-Iraqi H, Duskin-Bitan H, Shimon I, Gorshtein A. Outcomes of hyperglycemia in patients with and without diabetes hospitalized for infectious diseases. *Diabetes Metab Res Rev*. 2018;34:e3027.
8. Glynn N, Owens L, Bennett K, Healy ML, Silke B. Glucose as a risk predictor in acute medical emergency admissions. *Diabetes Res Clin Pract*. 2014;103:119-126.
9. Lipska KJ, Ross JS, Wang Y, et al. National trends in US hospital admissions for hyperglycemia and hypoglycemia among medicare beneficiaries, 1999 to 2011. *JAMA Intern Med*. 2014;174:1116-1124.
10. Report of a World Health Organization Consultation. Use of glycated haemoglobin (HbA1c) in the diagnosis of diabetes mellitus. *Diabetes Res Clin Pract*. 2011;93:299-309.
11. Manley SE, O'Brien KT, Quinlan D, et al. Can HbA1c detect undiagnosed diabetes in acute medical hospital admissions? *Diabetes Res Clin Pract*. 2016;115:106-114.
12. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag; 2016.
13. Meyer D, Zeileis A, Hornik K. The Strucplot framework: visualizing multi-way contingency tables with vcd. *J Stat Softw*. 2006;17:1-48.
14. Meyer D, Zeileis A, Hornik K. vcd: Visualizing Categorical Data. R package version 1.4-4. 2017.
15. James J, Kong M-F, Berrington R, Dhatariya K. Diabetes at the front door. A guideline from the Joint British Diabetes Society (JBDS) for Inpatient Care Group. https://abcd.care/joint-british-diabetes-societies-jbds-inpatient-care-group. Accessed March 10, 2020.
16. Report from house of Commons library. https://commonslibrary.parliament.uk/social-policy/health/diabetes-in-england-whereare-the-hotspots. Accessed March 10, 2020.
17. https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/demographics/age-groups/latest. Accessed March 09, 2020.
18. Webber J, Chua S, Cockwell P, et al. Effects of concurrent illnesses and treatments on surrogate glycaemic markers. *Diabet Med*. 2017;34(Suppl 1):P138.
19. Bhattacharjee D, Vracar S, Round RA, et al. Utility of HbA1c assessment in people with diabetes awaiting liver transplantation. *Diabet Med*. 2019;36:1444-1452.
20. Herman WH, Ma Y, Uwaifo G, et al. Differences in A1C by race and ethnicity among patients with impaired glucose tolerance in the Diabetes Prevention Program. *Diabetes Care*. 2007;30:2453-2457.
21. Cohen RM. A1C: does one size fit all? *Diabetes Care*. 2007;30:2756-2758.
22. *NICE Guidelines (PH38) Type 2 Diabetes: Prevention in People at High Risk*. London: NICE; 2012.
23. Manley S, Nightingale P, Stratton I, et al. Diagnosis of diabetes: HbA_{1c} versus WHO criteria. *Diabetes Prim Care*. 2010;12:87-96.
24. Levi OU, Webb F, Simmons D. Diabetes detection and communication among patients admitted through the Emergency Department of a Public Hospital. *Int J Environ Res Public Health*. 2020;17(3):980. https://doi.org/10.3390/ijerph17030980

How to cite this article: Ghosh S, Manley SE, Nightingale PG, et al. Prevalence of admission plasma glucose in 'diabetes' or 'at risk' ranges in hospital emergencies with no prior diagnosis of diabetes by gender, age and ethnicity. *Endocrinol Diab Metab*. 2020;3:e00140. https://doi.org/10.1002/edm2.140

9860–9873 *Nucleic Acids Research, 2017, Vol. 45, No. 17* *Published online 18 August 2017*
doi: 10.1093/nar/gkx714

Meta-analysis of transcriptomic datasets identifies genes enriched in the mammalian circadian pacemaker

Laurence A. Brown[1], John Williams[2], Lewis Taylor[1], Ross J. Thomson[1], Patrick M. Nolan[2], Russell G. Foster[1,*] and Stuart N. Peirson[1,*]

[1]Sleep and Circadian Neuroscience Institute (SCNi), Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, OX1 3RE, UK and [2]MRC Harwell Institute, Harwell Campus, Oxfordshire OX11 0RD, UK

Received September 07, 2016; Revised July 27, 2017; Editorial Decision July 30, 2017; Accepted August 04, 2017

ABSTRACT

The master circadian pacemaker in mammals is located in the suprachiasmatic nuclei (SCN) which regulate physiology and behaviour, as well as coordinating peripheral clocks throughout the body. Investigating the function of the SCN has often focused on the identification of rhythmically expressed genes. However, not all genes critical for SCN function are rhythmically expressed. An alternative strategy is to characterize those genes that are selectively enriched in the SCN. Here, we examined the transcriptome of the SCN and whole brain (WB) of mice using meta-analysis of publicly deposited data across a range of microarray platforms and RNA-Seq data. A total of 79 microarrays were used (24 SCN and 55 WB samples, 4 different microarray platforms), alongside 17 RNA-Seq data files (7 SCN and 10 WB). 31 684 MGI gene symbols had data for at least one platform. Meta-analysis using a random effects model for weighting individual effect sizes (derived from differential expression between relevant SCN and WB samples) reliably detected known SCN markers. SCN-enriched transcripts identified in this study provide novel insights into SCN function, including identifying genes which may play key roles in SCN physiology or provide SCN-specific drivers.

INTRODUCTION

Life on Earth has evolved under a predictably changing cycle of light and darkness and, as a result, virtually all organisms demonstrate striking changes in physiology and behaviour over the 24-h day. These rhythms are not simply a response to the changing environment, but persist under constant conditions—providing evidence for the existence of an endogenous biological clock. In mammals, the site of the master circadian pacemaker is the suprachiasmatic nuclei (SCN) in the anterior hypothalamus (1). The SCN receives light information from the retina via the retinohypothalamic tract, synchronizing (entraining) SCN rhythms to the external environment. SCN lesions result in loss of physiological and behavioural rhythms, and transplantation of foetal SCN can restore rhythmicity with a period consistent with that of the donor tissue (2). The circadian clock is the product of an intracellular transcriptional-translational feedback loop (TTFL), comprised of a number of so-called 'clock genes'. Whilst clock genes are expressed in tissues throughout the body, the coordinated function of the circadian system depends upon neural, hormonal and behavioural output of the SCN pacemaker (3–5).

The identification of mammalian clock genes has critically depended upon the application of forward and reverse genetics. Large scale mutagenesis projects have been invaluable in identifying novel mouse mutants with circadian phenotypes. Mapping the underlying mutation has led to the identification of a number of key clock genes, including *Clock* (6), *Fbxl3* (7), *Fbxl21* (8) and *Zfhx3* (9). By contrast, reverse genetics has been used to identify homologs of *Drosophila* clock genes (identified by forward genetics), including *Period 1–3* (10–12) and *Cryptochrome 1–2* (13,14) as well as proteins that interact with known clock components, such as *Arntl* (*Bmal1,* (15)). A key feature of many of the core clock genes is that they are rhythmically expressed over the 24-h cycle. Therefore, studies to identify new clock genes have often taken the approach of identifying transcripts that are rhythmically expressed in the SCN or other peripheral clocks, under either entrained light/dark conditions or over one or more circadian cycles under constant conditions (16). These studies have been successful in identifying novel clock genes as well as genes important for SCN function (17,18). However, not all clock genes are rhythmically ex-

*To whom correspondence should be addressed. Tel: +44 0 1865 618 674; Email: stuart.peirson@eye.ox.ac.uk
Correspondence may also be addressed to Russell G. Foster. Tel: +44 0 1865 618 661; Email: russell.foster@eye.ox.ac.uk

pressed, and rather than changing in abundance, some components of the TTFL may show rhythmic post-translational modification or simply depend upon interactions with other rhythmic components. Moreover, many genes involved in key functions of the SCN are not elements of the TTFL. For example, a number of neuropeptides and their receptors play key roles in SCN physiology (such as vasoactive intestinal polypeptide, VIP (19) and VIPR2 (20)), as do GABA receptors (21).

An alternative strategy to identify mechanisms critical for SCN function is to characterize those genes that are selectively enriched in the SCN. The most straightforward way of identifying SCN-enriched genes is to directly compare the transcriptome of SCN against the whole brain. Despite the value of this approach, only a few such studies exist - focusing on multiple brain regions (22), or specifically on a subset of genes (23). Another advantage of this approach is that it enables SCN-specific genes to be identified, providing critical tools for conditional transgenesis. As many clock genes are expressed throughout the body, studies of constitutive knockout mouse models can be problematic due to developmental effects or differing roles of these genes in other target tissues/organs. As a result, many researchers are looking to the use of conditional knockouts, for example, using Cre-lox technology (24,25). The identification of genes that are specifically and highly enriched in the SCN may not only reveal new biological insights regarding SCN physiology, but also has the potential to produce SCN-specific drivers, which would be of benefit to the circadian community.

Studies using transcriptomic approaches face several well-characterized challenges (26). Biological and technical variance result in a degree of noise in any study, but when making comparisons between thousands of genes, false positive rates become a major problem. Using $P < 0.05$, comparing the expression of ∼20 000 transcripts would be expected to give 1000 false positives—genes that would appear significant even though they are unchanged. As such, false discovery rates are corrected to account for this issue. However, due to the cost of running transcriptomic experiments, sample sizes are often limited. This results in reduced power—that is the ability to identify real differences where they exist, which can result in biologically relevant findings being missed. As such, transcriptomic studies are often a statistical balancing act—resulting in a list of candidate genes which may contain false positives, and may be missing real genes of interest. The downstream effects of this compromise will often be that examining pathways or gene ontologies may be uninformative or misleading.

One way of addressing these statistical issues is to gather as much relevant data as possible from all available sources. Such approaches are widely applied to clinical studies in the form of meta-analysis—a combination of all available data in the literature to enable better outcome decisions, which are now considered essential in clinical science (27). Published microarray data are widely available via deposition into public databases, including NCBI's Gene Expression Omnibus (28) and EBI's ArrayExpress (29). As such, meta-analysis of microarray data is a topic of growing interest. This has led to a series of possible methods for dealing with the problems inherent in comparing different microarray platforms with varying numbers of probes for different numbers of transcripts. The methods for drawing together a series of experiments with differing biological and technical variance are numerous, and depend partly on the quality and detail of the available data. Where only lists of significant genes are available, methods such as vote-counting provide the most straightforward approach to meta-analysis (30). Venn-diagrams are often used as a simple form of vote counting, focusing on genes that pass both a fold-change criteria and one or more statistical tests. A problem of this approach is that the pre-processing methods and inclusion criteria used by the original authors to construct each gene list are unlikely to be comparable across all data sources (16). Alternative approaches make use of original raw data by combining either rank-orders of genes (31) or *P*-values as measures of significance (32). Finally, weighting the contribution of each data point for a gene by the inverse of the variance of that data provides a way of comparing data across studies, assuming that lower variance is an indicator of reproducibility (33). These methods are discussed in detail by Ramasamy *et al.* (34), which provides the basis for our analysis.

Here, we describe a meta-analysis of expression data from SCN and whole brain across a range of microarray technologies and incorporating recent RNA-Seq data (Figure 1). We use the term 'platform' in the rest of the paper to refer to each different type of microarray technology (e.g. Mouse Exon ST1.0 arrays) or RNA-Sequencing. This meta-analysis enables us to reliably identify transcripts enriched in the SCN. These transcripts provide novel insights into the function of the master circadian pacemaker, providing new candidates that may play key roles in SCN function, as well as providing new targets for SCN-specific gene targeting.

MATERIALS AND METHODS

Data-mining to identify SCN and WB datasets for Affymetrix platforms

Analysis of publicly available datasets revealed that a large part of the apparent variation in any comparison was attributable to differences in the different array platforms. For this reason samples from a particular microarray platform could only be used if both SCN and whole brain (WB) data were available for that platform. In addition, investigations were limited to the Affymetrix microarray platforms (by far the most prominent platforms with suitable data) and we only used data where the original array scan files (.CEL files) were available. This has previously been referred to as feature level extraction output (FLEO), and allows a higher level of control over the processing of array files and consequently the accuracy of comparisons at the gene level (34).

Datasets were identified using simple searches of depositories of publicly available microarray data: NCBI's Gene Expression Omnibus (28) or the EBI ArrayExpress (29). Search terms included 'suprachiasmatic', 'SCN', 'adult', 'wild-type' and 'whole brain'. Only using data where MIAME (Minimum Information About a Microarray Experiment, https://www.ncbi.nlm.nih.gov/geo/info/MIAME.html), has been supplied and where FLEO is possible (i.e. those with .CEL files available). Whole brain data gathered usually consisted of the control (or sham) samples without/before treatment, but not those within treatment

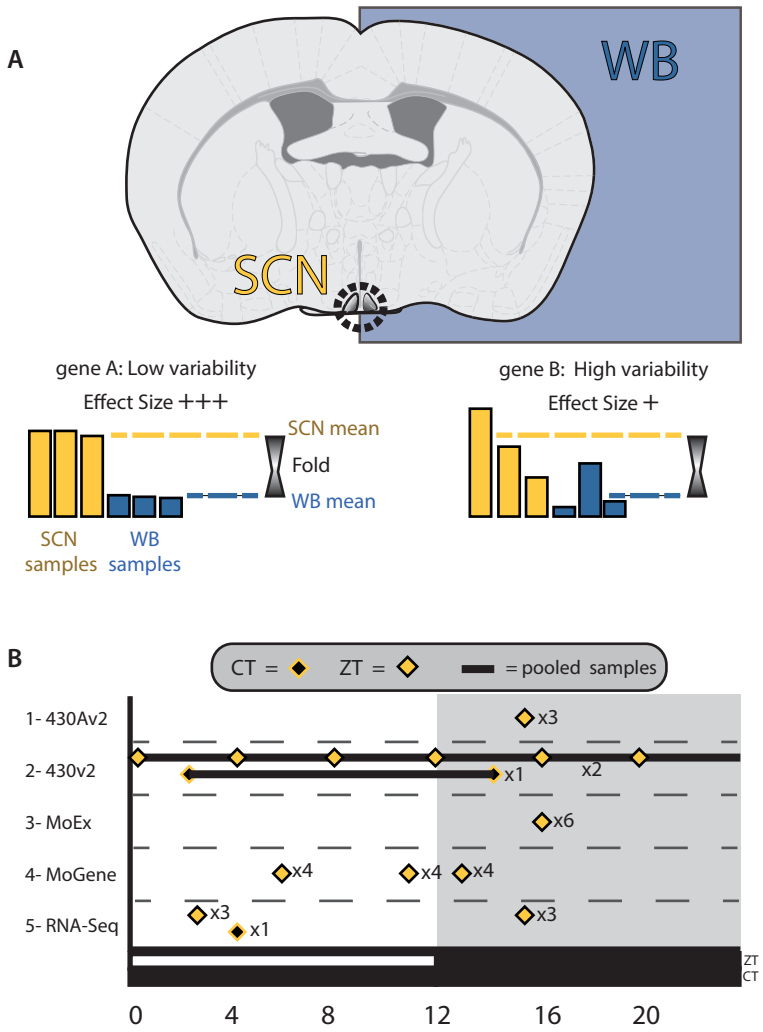


Figure 1. (**A**) The rationale for meta-analyses, with differences in expression between SCN and whole brain samples used to calculate effect sizes for each gene and transcriptomic platform. Treating each platform as an individual study allows data from disparate sources to be brought together in a meta-analysis. (**B**) The times at which SCN samples were collected, showing a range of protocols, including collection by external/zeitgeber times (ZTs) and internal/circadian times(CTs), as well as individual samples at multiple time points and the pooling of RNA samples prior to measurements of transcripts.

groups. Likewise, SCN data were restricted to sham treatments and wild-type animals. The majority of the whole brain data reported the inclusion of the cerebellum and olfactory bulbs in samples. We limited the comparison to adult mice, but with no restriction on strain or sex. The microarray data used in the current study are detailed in Table 1 (a more detailed in version of this table can be found in Supplementary File 1).

Feature level extraction output (FLEO)

Microarray data were processed as series of comparisons for each microarray platform, or RNA-Seq data. CEL files were imported in to AltAnalyze (version 2.9.0.2) and processed via integration with Affymetrix power-tools (APT), using the Robust Multi-array Averaging (RMA-sketch) algorithm. Following removal of probes known to cross-hybridize, the remaining probe-sets were then mapped to build 72 of the ENSMBL mouse database (GRCm38.p1). For exon-arrays (Affymetrix MoEx 1.0ST) expression at the gene level was examined, with this achieved using only those probe-sets that match to exons present in all expressed known splice variants of the ENSMBL transcript in question (constitutive probe-sets). For the comparisons of 3′ arrays (430v2 and 430Av2), those probe-sets known to cross-hybridize (_x_ and _s_ -appended) and those without mappings being excluded. The average values of Hedges' g and Variance were calculated for each gene symbol or ENSMBL GID.

Microarray pre-processing for meta-analysis

Effect sizes (ES) for each comparison were calculated as Hedges' g values (35). Briefly, this involves calculation of Cohen's d value (log2 fold-enrichment SCN vs WB, divided by pooled standard deviation), followed by an adjustment

Table 1. Data collected for the meta-analysis as a series of comparisons (studies) within Affymetrix microarray platforms and RNA-Seq data

Accession code	File codes	SCN or WB	PMID	Arrays/ Samples	Study: Platform
GSE6904	GSM159292–4	SCN	18021443	3 SCN vs 8WB	1: 430AV2 Mouse
GSE7814	GSM189596/598/600/602	WB	17991715		
GSE7814	GSM189628/630/632/634	WB	17991715		
GSE16496	GSM414571–2	SCN	21858037	3 SCN vs 13WB	2: 430V2 Mouse
GSE28574	GSM707557	SCN	21610730		
GSE20411	GSM511616–20	WB	20526689		
GSE20411	GSM511621–25	WB	20526689		
GSE9954	GSM252077–9	WB	18365009		
MEXP-3933	WT_sham_1–6	SCN	23993098	6 SCN vs 22WB	3: MoEx 1.0ST
GSE27282	GSM674605–14	WB	21625610		
GSE27282	GSM674615–26	WB	21625610		
E-MEXP-3493	WT_ZT6(1to4), WT_ZT11(1to4), WT_ZT13(1to4)	SCN	22264613	12 SCN vs 12WB	4: MoGene 1.0ST
GSE34469	GSM849761–2, GSM933084–5	WB	23580197		
GSE34305	GSM847026–7/29/31/33	WB	22560501		
GSE24940	GSM613009–11	WB	21088282		
PRJEB9284	SAMEA3368226-SAMEA3368230, SAMEA3368237	SCN	26232227	7 SCN vs 10WB	5: RNA-Seq
PRJNA235222	SAMN02585109	SCN	24531307		
GSE43013	GSM1055111	WB	25677554		
GSE30352	GSM752614	WB	22012392		
GSE30352	GSM752615	WB	22012392		
PRJEB2494	SAMEA811980	WB	PMC3428933		
PRJEB2494	SAMEA811978	WB	PMC3428933		
GSE41338	GSM1015150/51	WB	23258891		
GSE41637	GSM1020640	WB	23258890		
GSE41637	GSM1020649	WB	23258890		
GSE41637	GSM1020657	WB	23258890		

The original experiments from which the data can be tracked by database codes or Pubmed IDs (references 91–103). A more detailed and extended version of this table can be found in the **Supplementary File 1**.

of number of arrays (known as the j factor). The variance of each ES was also calculated.

RNA-Seq pre-processing for meta-analysis

Published RNA-Seq projects were obtained from the Sequence Read Archive (Table 1, and in more detail in Supplementary File 1). Sequence reads were aligned to the mm10 genome using TopHat2 (36). When unpublished, library strand direction was confirmed with manual inspection in the Integrative Genomics Viewer (37). Junction BED files were sent to AltAnalyze for downstream gene annotation, counting and differential expression analysis (38). Per-identifier ES values were then calculated as described above.

Meta-analysis

All data were indexed against MGI gene symbols (using BioMart mouse version 72), with a total of 31 684 MGI gene symbols having data for at least one platform. The combined effect size was calculated as described for the inverse variance methods described by Choi *et al.* (33). We calculated effect sizes and significance for the inverse variance meta-analysis based on a random effects model (REM), as this does not assume that there is a single common effect size, but rather a range of true effect sizes with additional sources or variation. Basic calculation and indexing was carried out from the output of AltAnalyze for each study. The analysis was carried out using the PyData stack in the Python programming language (version 2.7.11, as part of the Anaconda python distribution version 4.0.0, from https://www.continuum.io/downloads and is available as a series of interactive notebooks (at https://github.com/LozRiviera/SCN_enrich_Meta, DOI: 10.5281/zenodo.324907). From the meta-analysis, *Z*-values were used to calculate *P*-values from two-tailed tests and subsequently to apply a multiple-testing correction in the form of the false discovery rate (FDR) *q*-value (39). The subsequent FDR-adjusted *q*-values were calculated from *P*-values in R (version 3.2.2, (40)), using the 'qvalue' package (version 2.0) (41).

Resources to examine the distribution of SCN-enriched transcripts

Confirmation of potential SCN-enriched genes using online resources involved searches by gene symbol on both the Allen Brain Atlas (42) and the GENSAT database (43) .

Immunohistochemistry

All work was carried out in accordance with Animal [Scientific Procedures] Act 1986, with procedures reviewed by the clinical medicine animal care and ethical review body (AWERB) for the University of Oxford, and conducted under project licence PPL 30/2812 and personal licence IDB24291F. Young-adult (8–24 weeks of age) male wild-type C57BL/6J mice (RRID:IMSR_JAX:000664), were obtained from Envigo (Alconbury UK) and housed in specific pathogen free conditions, with the only reported positives on health screening over the entire time course of these studies being for *Helicobacter hepaticus* and *Entamoeba spp.* All

animals were singly-housed, provided with food and water *ad-libitum* and maintained on a 12-h light:12-h dark cycle (150–200 lux, cool white LED, measured at the cage floor), in light-tight environmental enclosures.

Brains were fixed by perfusion with (then immersion overnight in) 4% paraformaldehyde in phosphate-buffered saline (PBS). Following cryoprotection in 30% sucrose in PBS, brains were embedded in optimal cutting temperature compound (OCT, VWR International Ltd.) and sectioned at a thickness of 14–20 μm. The primary antibody for SYTL4 (1:200–1:500, Rabbit polyclonal, ab110519, Abcam plc, Cambridge, UK, RRID:AB_10858160) was used following blocking with 10% Donkey serum in PBS + 0.1% Triton X-100 + 0.1% Tween 20 (serum and detergents from Sigma-Aldritch Ltd., Dorset, UK). Validation of the primary antibody is included in Supplementary File 1. Detergents were excluded following the blocking step to prevent loss of lipid micro-domains due to excessive permeablization. The secondary antibody was Donkey anti-rabbit Alexa Fluor® 568 conjugate (1:200, Thermo-Fisher Scientific Inc.). Confocal images were collected using a Zeiss LSM710 (Carl Zeiss Ltd., UK).

Gene ontology (GO) and network analysis

Data were imported into Cytoscape (44) (version 3.3.0). Established protein–protein interactions were obtained from the STRING database (45) (http://string-db.org), using the list of both enriched and depleted genes (gene symbols) and the StringApp plugin for Cytoscape (version 0.9.2). The BiNGO plugin (version 3.0.3) (46) was used to show gene ontology (GO) terms that are over-represented when compared by Hypergeometric test to the whole GO annotation, following Benjamini–Hochberg false discovery rate (α = 0.05). GO (47) files and MGI annotations were obtained as of 4 February 2016 and the tests were run using gene symbols. Visualization of the results of the ontology analyses were carried out using the Enrichment Map plugin for Cytoscape (version 2.1.0) (48), with defaults for BiNGO files (p-cutoff = 0.001, q-cutoff = 0.05, Jaccard similarity cutoff = 0.25).

RESULTS

Identification of SCN-enriched and SCN-depleted transcripts

Restricting data to MIAME-compliant datasets where Affymetrix CEL. files were deposited, a total of 79 microarrays were obtained (24 SCN and 55 WB), from four different microarray platforms, alongside 17 RNA-Seq data files (7 SCN and 10 WB). These data are summarized in Table 1 (and in more detail in Supplementary File 1). Indexing to MGI gene symbols resulted in 31 684 symbols having data for at least one platform, with the number of symbols covered by each platform and the degree of overlap between platforms shown in Supplementary File 1 and Table 1.

Available data were processed as five studies based on the microarray platform (or RNA-Seq). The meta-analysis was conducted using an REM for weighting individual effects scores. These scores were themselves derived from the differences in expression between the SCN and WB samples in each study ($\log_2$-transformed fold-changes). The data are summarized in Figure 2 and is also implemented in an interactive version to improve the accessibility of the data (Supplementary File 2). All MGI gene symbols with at least one datum (study) can be found in the Supplementary File 3. Constraining the results using a positive false discovery rate (pFDR) q-value of 0.01 (expected that around 1% of results are false-positives) gives a list of 4403 symbols. Where discussed in the remainder of the paper, only genes passing this pFDR correction will be referred to as 'enriched' or 'depleted'. To obtain a manageable list of genes for pathway analysis these 4403 genes were further restricted to those with and a combined effect size (M*) of 3 or more in either direction (Figure 2), leaving a list of 1037 gene symbols. Of these, 426 were enriched and 611 depleted.

The top 20 enriched gene symbols from meta-analysis (ranked by combined effect size) are presented in Table 2, with corresponding evidence from online resources for gene expression. http://www.alleninstitute.org (42) and http://www.gensat.org/ (43). It is notable that many of the genes associated with the known function of the SCN, such as canonical clock genes, are not present in this list. This is in part due to the nature of the study, with both expression in other nuclei in the brain and variations in expression over time (rhythmic or otherwise) being likely to increase variation for such a transcript, therefore decreasing its enrichment score. It is the case that many genes such as *Vip* show combined effect sizes well above zero (see Table 3, a table of genes with established roles in the SCN), but are excluded from the current list of 4403 genes by q-value.

Validation of an SCN-enriched transcript, Sytl4

It is possible that the combined effect size alone may be a good indicator of genes with important roles in the function of the SCN. To examine this further, we selected another gene with a high M* value (8.35, ranked fourth by M* in list of 4403 gene symbols at 1% pFDR) that has not previously been described in the SCN, *Synaptotagmin-like 4* (*Sytl4*, or *Granuphilin-a*). The protein is known to be involved in the controlling of the release of dense-core vesicles and exosomes, but has no known role in the SCN to date. Immunohistochemistry on fixed SCN tissue revealed that the SYTL4 was indeed expressed in the SCN, as well surrounding hypothalamic nuclei and was largely absent from the rest of the brain (see Figure 3, an example of the distribution of SYTL4, from tissue collected at ZT18). The main known binding partner of SYTL4, RAB27A, is also enriched in the SCN (Rab27a, M* = 3.40) (49).

Pathway analysis

To further examine the potential relationships between the 1037 transcripts, both enrichment of GO terms and known interactions were examined. Figure 4 shows known interactions (of high confidence) between those gene symbols in the list of 1037, as revealed by the STRING database. Like much of the brain, the SCN relies on a balance of both glutamatergic and GABAergic neurotransmission and as such the levels of some transporters for glutamate and GABA differ greatly. The vesicular GABA transporter VGAT (encoded by *Slc32a1*) is consistently enriched throughout all

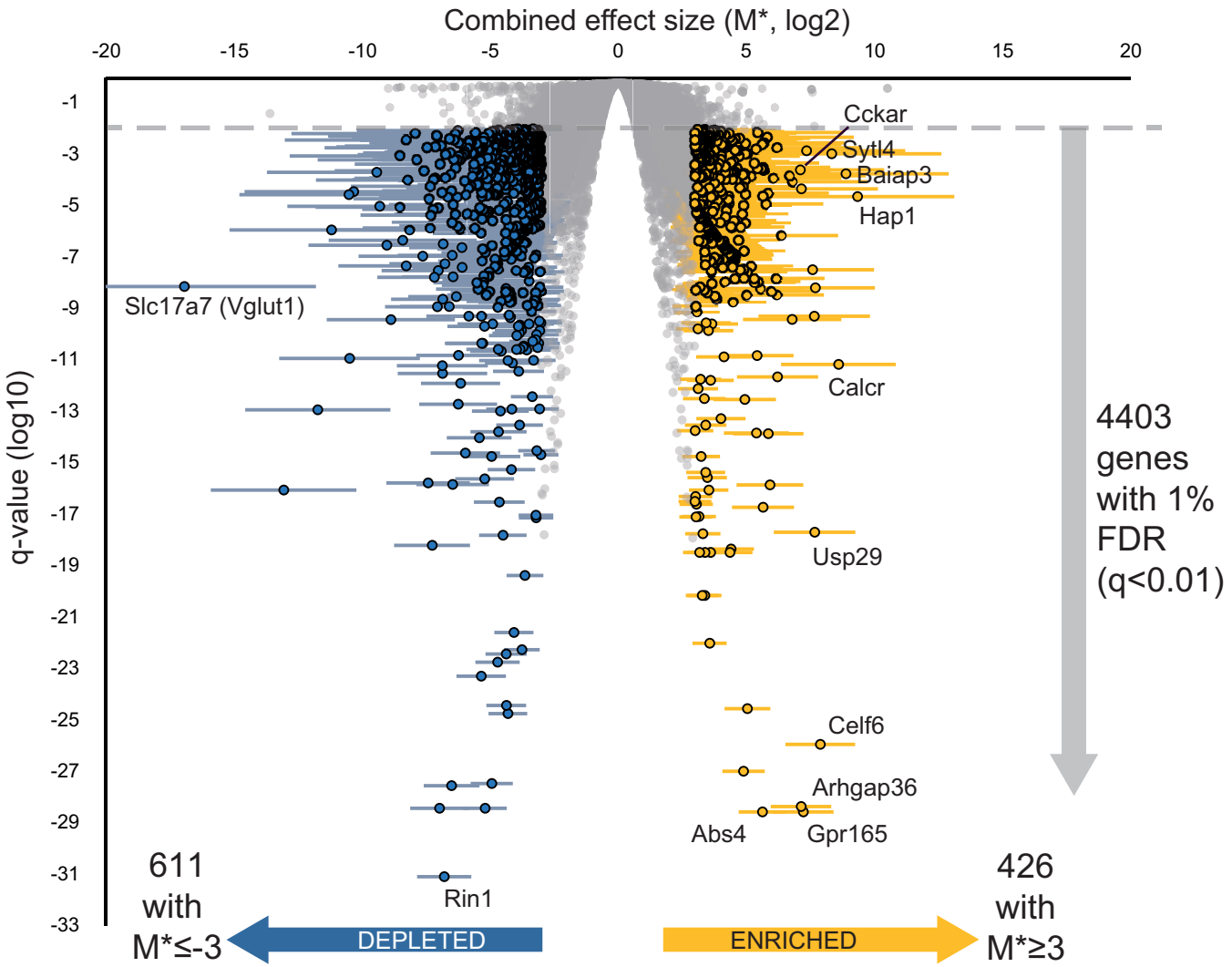


Figure 2. Volcano plot of genes found to be enriched in the SCN by meta-analysis. Plot shows enrichment in SCN versus WB (combined effect size, M*, derived from $\log_2$-transformed fold changes) against the significance (FDR-corrected q-value), with error bars showing 95% confidence intervals. Genes with $q < 0.01$ and M* ≥ 3 shown in gold (426 highly enriched) and those with $q < 0.01$ and M* ≤ -3 shown in blue (611 highly depleted). Genes labelled are the top enriched genes by combined effect size, or by significance, as well as the most depleted and most significant depleted genes.

the studies in the meta-analysis (M* = 3.41), whereas the vesicular glutamate transporter VGLUT1 (*Slc17a7*) is the most depleted transcript in the SCN when compared to the brain on the whole (M* = −16.93). Peptidergic transmission is also known to play a vital role in the SCN and other hypothalamic nuclei, (Figure 4, **box A**.), with the enrichment of GO terms suggesting that the axoneme and primary cilium, a focal point for GPCR-mediated signalling (50), may also be important for communication between cells in the SCN (Figure 5). Where families of genes/proteins are known to be important in the tissue of interest, a list of enriched transcripts may help identify the most relevant isoforms for further research. In the case of adenylate cyclases it seems that the SCN shows enriched expression of isoforms 6 and 7 and substantial depletion of isoforms 1 and 9 in particular (Figure 4, **box B**.). One clear finding from pathway and ontological analysis is that substantial parts of the synapse and many voltage-gated channels are depleted in the SCN. This is shown in the interactions drawn from the STRING database for K^+ channels (Figure 4, **box C**) and throughout the enrichment of GO terms related to both axons and dendrites (Figure 5). Additionally, whilst the neurotransmitter GABA and its receptors are ubiquitously expressed throughout the CNS, in the SCN the expression of subunits that comprise GABA-A receptors suggest that the stoichiometry may differ from much of the rest of the brain (Figure 4, **box D**). The complete lists of enriched GO terms is provided as Supplementary File 4.

DISCUSSION

The SCN provides a fascinating interplay between intracellular molecular clocks and the synchronized daily rhythms of neuronal activity (4,51). Although this rhythmic control over neuronal activity continues in the absence of external influences, the circadian pacemaker must be capable of responding to photic cues as well as inputs from other brain

Table 2. Top 20 results from meta-analysis (ranked by combined effect size), with corresponding evidence from online resources for gene expression

MGI gene symbol	Description	Number of studies	Combined effect size (M*)	pFDR *q*-value	Allen Brain Atlas	GENSAT
Hap1	huntingtin-associated protein 1	5	9.36	2.42E-05	14890	91396
Baiap3	BAI1-associated protein 3	5	8.90	1.87E-04	75081206	85221
Calcr	calcitonin receptor	5	8.61	7.17E-12	12096	60456
Sytl4	synaptotagmin-like 4	5	8.35	1.13E-03	75651223	x
Celf6	CUGBP, Elav-like family member 6	4	7.91	1.18E-26	71358616	66740
Ahi1	Abelson helper integration site 1	5	7.71	6.84E-09	69549642	x
Usp29	ubiquitin specific peptidase 29	2	7.68	2.14E-18	70194636	x
Ngb	neuroglobin	5	7.67	5.38E-10	79556712	x
Zcchc12	zinc finger, CCHC domain containing 12	4	7.60	3.42E-08	73817424	80305
Vwa5b1	von Willebrand factor A domain containing 5B1	3	7.36	1.47E-03	51559	x
Arhgap36	Rho GTPase activating protein 36	4	7.24	2.80E-29	69352834	x
Gpld1	glycosylphosphatidylinositol specific phospholipase D1	5	7.16	4.88E-05	74509585	x
Gpr165	G protein-coupled receptor 165	4	7.15	4.54E-29	70560278	75666
Cckar	cholecystokinin A receptor	5	7.12	2.62E-04	203	x
Tmem130	transmembrane protein 130	4	6.80	4.03E-10	89067	x
Chodl	chondrolectin	5	6.80	9.14E-05	71380977	x
Itih3	inter-alpha trypsin inhibitor, heavy chain 3	5	6.69	1.58E-04	600	x
Nap1l5	nucleosome assembly protein 1-like 5	5	6.38	7.34E-07	72080123	x
Scn9a	sodium channel, voltage-gated, type IX, alpha	5	6.23	2.34E-12	71325438	x
RP24–361E14.1	x	1	6.20	3.62E-09	x	x

http://www.alleninstitute.org and http://www.gensat.org/. 'x' indicates no data or identifier currently available.

Table 3. The enrichment of previously suggested SCN markers and genes important for signalling or rhythmicity, along with the combined effect size (M*) derived from the meta-analysis

MGI gene symbol	Effect size (M*)	Pubmed ID
Adcyap1 (PACAP)	2.20	9065523
Avp	**4.93**	25741730
Drd1a	−2.32	25643294
Lhx1	2.52	21525287
Nms	**3.86**	15635449
Prok2	**3.19**	12024206
Rgs16	**3.11**	21610730
Scg2	**4.70**	17319750
Six3	**3.85**	21525287
Six6	**5.67**	21525287
Syt10	**3.29**	21921292
Vip	6.12	11207820
Vipr2	**2.48**	12086606

Values in bold are those that meet the relevant cut-off (q-value < 0.01). Pubmed IDs provided for examples of the many references suggesting functional involvement in the SCN.

regions. These processes may involve a different set of genes and proteins, not all of which are expected to be rhythmically expressed. The current study helps to identify those transcripts, rhythmic or otherwise, that are found in the SCN more often that in the brain as a whole. We have shown that 4403 genes are expressed at a significantly higher (enriched, 2346 with $q < 0.01$) or lower level (depleted, 2057), when compared to the brain as a whole. Furthermore, the approach taken uses only existing public data and would continue to grow in strength as more data are deposited in public repositories.

This meta-analysis has provided a robust list of transcripts that are differentially expressed in the SCN when compared to the rest of the brain. This includes transcripts which are selectively enriched in the SCN which provide new avenues for understanding SCN physiology. Amongst the transcripts found to be selectively enriched in the SCN are genes already known to play a role in SCN function or circadian timing, or transcripts that have been used as markers of the SCN (Table 3). These transcripts include the genes for well-established peptide neurotransmitters such as arginine-vasopressin (*Avp*), Neuromedin S (*Nms*) and Prokineticin 2 (*Prok2*) (52–54). In addition to signalling molecules, genes such as *Ngb*, and *Zfhx3* are highly expressed in the SCN. Knockout of Neuroglobin (*Ngb*, M* = 7.67), increases responses of the SCN to light-pulses via the retina (55) and *Zfhx3* is a zinc-finger homeobox domain gene, recently shown to act via AT motifs on many neuropeptide promoters (9).

The focus of the current study was to look at over-represented transcripts in the SCN, rather than rhythmic ones. As summarized in Figure 1B, SCN samples were taken from a range of times (both ZT and CT), with some of the samples pooled prior to measuring transcription, there-

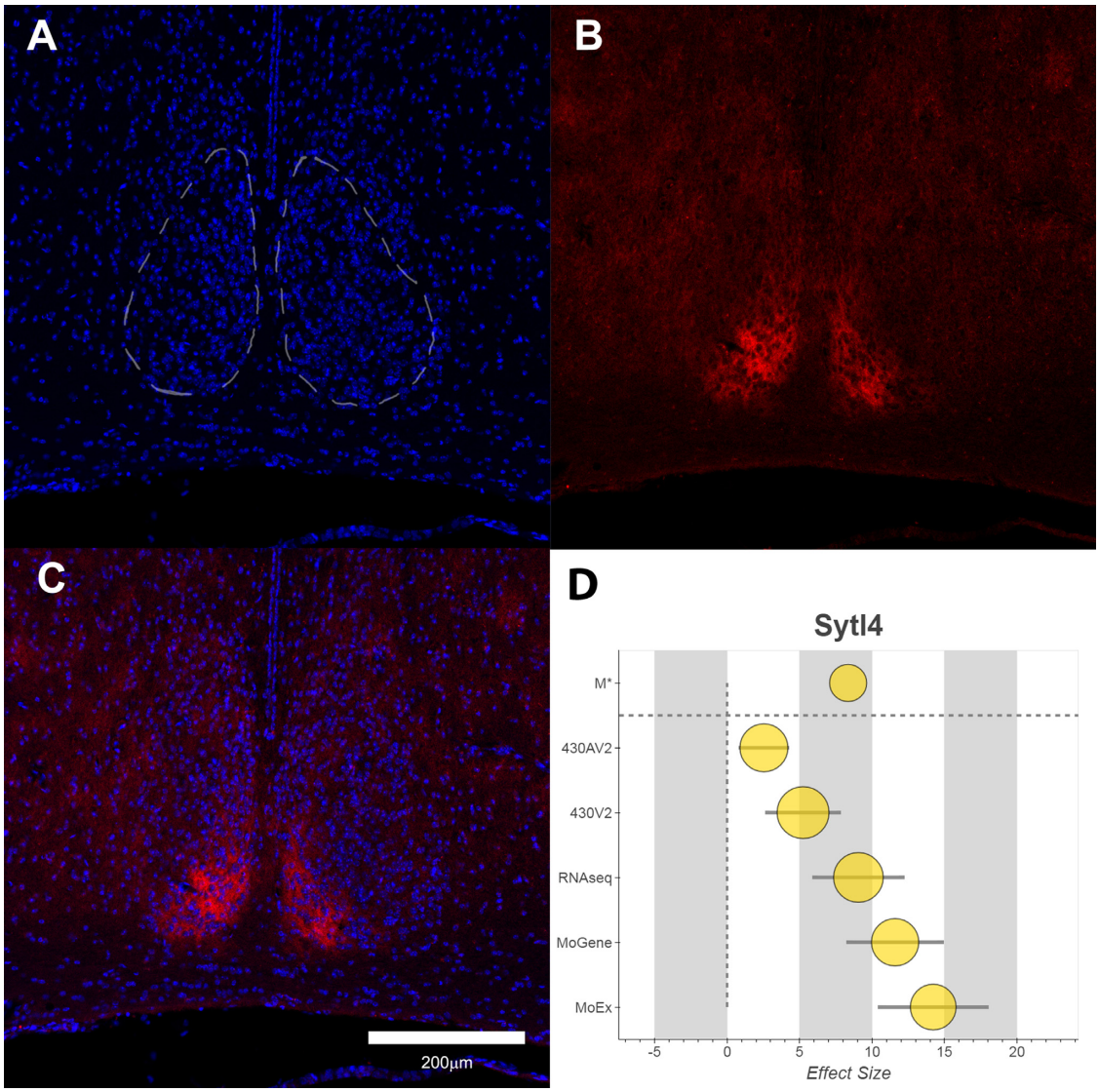


Figure 3. Distribution of the SYTL4 protein in the Hypothalamus. (**A**) DAPI staining of Nuclei shown nuclear-dense SCN and surrounding tissue. (**B**) SYTL4 shows strong expression in the dorsomedial SCN. (**C**) Merge of both channels. (**D**) Forest plot of the data in the meta-analysis for *Sytl4*, consisting of the combined effect size (M*, top), followed by individual Hedges' *g* values for each platform/study. Error bars show variation within each study and node size represents the weighting towards the final M* value. The tissue in this figure was from a C57BL/6J male mouse, collected at ZT18.

fore making it difficult to further assess rhythmicity. Indeed, the methodology may select against those genes that cycle in expression as this will increase the variation in these data. An example is the vasoactive intestinal polypeptide gene (*Vip*), a neuropeptide transmitter with an established role in the SCN(19). Despite a high combined effect size (M* = 6.12), *Vip* is excluded from the list of enriched transcripts due to high variability (FDR q-value > 0.01), as is Magel2 (M* = 7.67). Mice lacking *Magel2* show normal rhythms and entrainment, but with a reduced amplitude and increased daytime activity and MAGEL2 is expressed predominantly within AVP-containing neurons in the SCN (56,57). There may be further genes with relevance to the SCN that show high M* values, but that are excluded due to their variability. Furthermore, with this selection against variable transcripts in mind, along with the fact that the core 'transcription-translation feedback loop' (TTFL) is expressed in cells throughout the body, it is not surprising that elements of the TTFL are not enriched in the current study (none of the 4403 transcripts with a $q < 0.01$). Some genes with roles in circadian rhythms are depleted (*Cry2*: M* = −3.30, *Npas2*: M* = −2.95, *Bhlhe40 (Dec2)*: M* = −2.89) when compared to expression in the whole brain.

Although the primary interest of this study is enriched transcripts, the analysis also identified transcripts that appear to be selectively depleted in the SCN. One notable finding is that large numbers of immediate early genes (IEGs) are found to be highly depleted in the SCN (Supplementary File 1, Supplementary Table 2). IEGs are important for converting experience to changes in plasticity in the brain, including the transcriptional responses to events such as light-pulses (58–60). Our findings might suggest that expression of IEGs is not required for circadian rhythmicity in the SCN and that the activation of IEGs from a low baseline of ex-

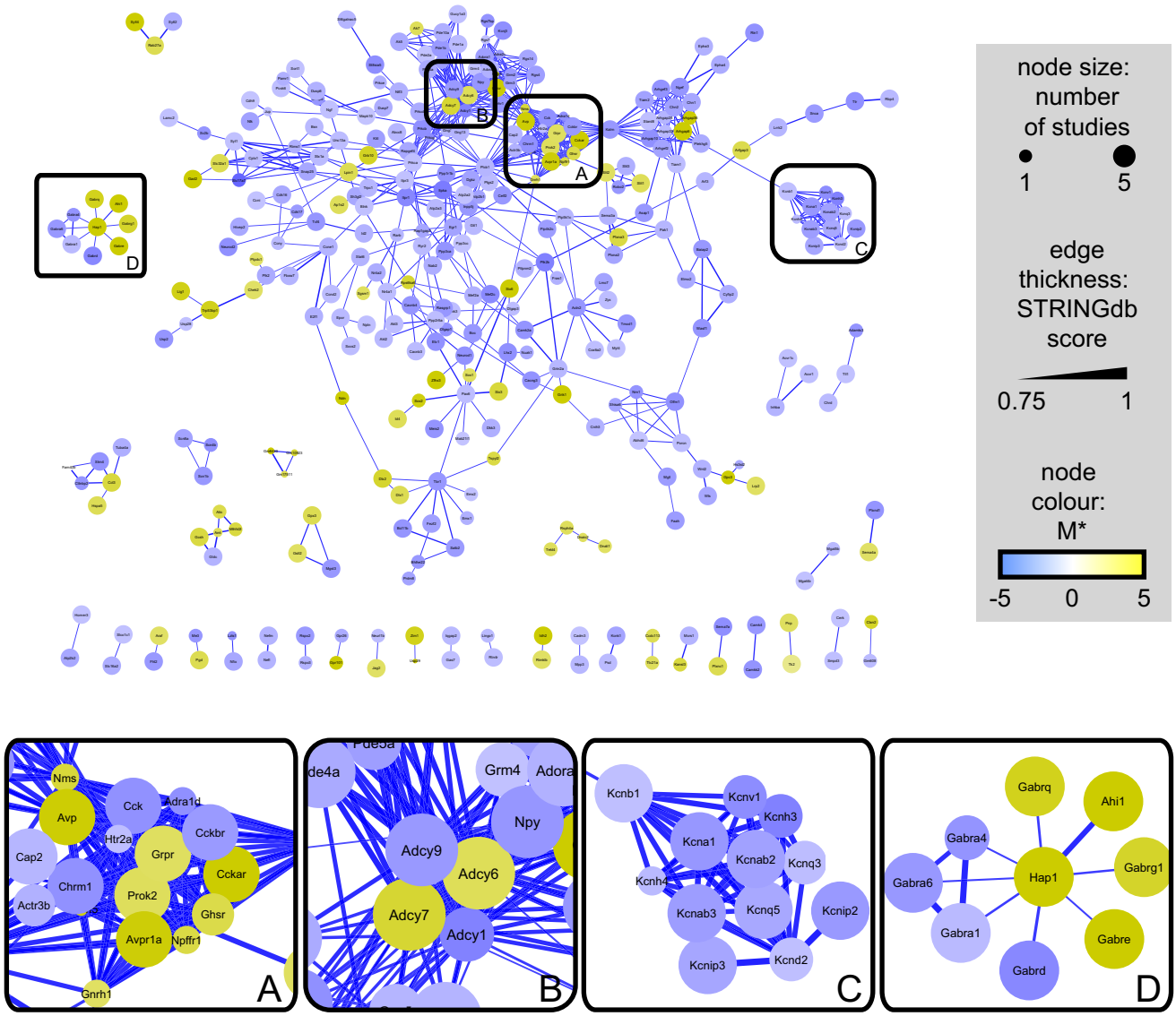


Figure 4. Network graph of SCN-enriched and SCN-depleted transcripts, constructed using interactions from the STRING database. Cut-off for interactions was an overall STRING score ≥0.75 and only those gene symbols with one or more interactions are displayed (339 of 1037). Side Panels (**A**–**D**) are enlargements of clusters within the network. Nodes are shown in gold for enriched transcripts and blue for depleted ones, with edge-thickness mapped to the overall score of the interaction from the STRING database (thickest lines = highest confidence).

pression may be vital for external stimuli to modify phase and plasticity in the SCN.

Intra-neuronal communication and synchrony in the SCN

The SCN is comprised of a network of cellular oscillators and the balance between synchronising and phase-repulsive signalling is essential for coordinated output from the circadian pacemaker (61,62). The density of synaptic connections (synapses per cell) has been reported as sparse (21,63) and the densely-packed neurons of the SCN possess smaller cytoplasmic compartments and reduced axonal and dendritic arborizations when compared to other neurons throughout the brain (64). These reports are in accordance with the overrepresentation of ontological terms such as 'synapse' and 'dendrite' within depleted transcripts (Figure 5 and Supplementary File 4) supporting this lack of extensive neuritic morphology in the SCN. The list of enriched genes contains many signalling molecules and their receptors that are already known play roles in coupling the cells of the SCN together (see Table 3). Proteomic analysis has previously pointed to the cycling of synaptic vesicles as essential for robust circadian synchrony (65). The importance of dense-core vesicles in a functional SCN is also indicated by the finding that double-knockouts for the transmembrane proteins of these vesicles, *Ptprn* (ia-2) and *Ptprn2* (ia-2β), show a loss of diurnal rhythms in heart-rate, temperature and locomotor activity (66). Many of the SCN-enriched transcripts identified in this study have no known role in the circadian pacemaker. One such gene, *Sytl4*, has been shown to play an important role in regulated exocytosis, controlling the release of insulin and other substances from dense-

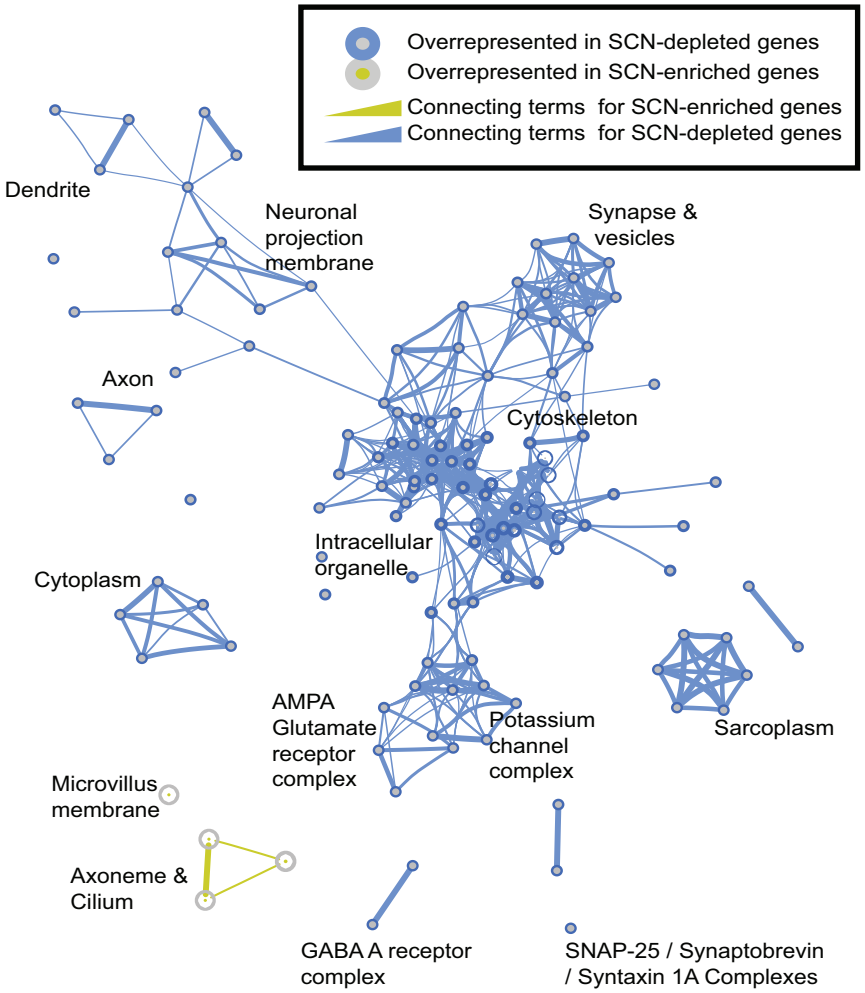


Figure 5. Enrichment map of the gene ontology data for SCN-enriched and SCN-depleted transcripts. Enrichment map generated from BiNGO output file for cell components (output from BiNGO analysis provided in Supplementary File 4).

core vesicles in tissues and cell lines (67,68). Other studies have reported hypothalamic expression and a role in facilitating sexually dimorphic behaviours in the hypothalamus (69). Immunohistochemistry in the current study confirms the enrichment of SYTL4 in the SCN, supporting the role of dense-core vesicles in SCN function.

In addition to the peptidergic communication, GABA-A receptor mediated signalling is known to be important to the functioning of the SCN, in particular the ability of GABA to promote both synchrony and disorder (21,62,70). Unlike many other areas of the brain, in the SCN it seems that some of the less-studied GABA subunits (*Gabre* and *Gabrq*) may play important roles, alongside the expression of multiple GABA transporters. This analysis also suggests that subunits of the recognized 'extra-synaptic' receptors (e.g. *Gabrd*, M* = −8.87) are selectively depleted in the SCN. Genes relevant to the function of GABAergic signalling in the SCN also extend to known interactors such as *Hap1*, which has been shown to control recycling of GABA-A receptors to the synapse, preventing internalization and recycling (71). *Hap1* (Huntingtin-associated protein 1) was discovered as a binding partner for Huntingtin and is therefore implicated in the neurodegenerative disorder Huntington's disease (72). These findings are of particular interest considering the circadian abnormalities reported in both human and animal models of this disease (73) and may be an important part of understanding GABAergic signalling, both from the SCN and within the SCN, where release of GABA can be either brief, or tonic over many hours (61).

Imprinting and SCN function

A number of the transcripts that show enrichment in the SCN are known to be imprinted (allelic bias towards either the maternal or paternal copy of the gene in question). For example imprinted genes such as the E3-ubiquitin ligase, *Ube3a* and the G-protein *Gnas* have effects on the stability of circadian rhythms and sleep architecture (74–76). Although these genes are not present in the list of 4403 enriched and depleted genes, when a list of 45 confirmed imprinted genes (77) was matched to these 4403 genes in our study (43 symbols matched) 13 of these genes were present (see Supplementary File 1, Supplementary Table 3). This is significantly more than would be expected by chance ($P = 0.0044$, 13 or more from 43 being picked in 4403 from a to-

tal pool of 31 684, hypergeometric distribution). The presence of 12 imprinted genes in the list of those substantially and significantly enriched in the SCN (of the 426 genes with M* >3 and $q < 0.01$) is even more significant ($P = 2.2e-11$, 12 or more from 43 being picked in 426 from a total pool of 31 684, hypergeometric distribution). Although the true extent of imprinting is still the subject of some discussion, recent papers have reported that parental allelic bias during development is often accompanied with a higher overall expression (78,79). These findings raise the interesting possibility that parent of origin effects may occur in the molecular mechanisms underlying mammalian circadian rhythms.

New pathways and implications

Recently, a number of consortia have released the first genome-wide association studies concerning circadian and sleep profiles (80,81). Genes common to both these studies include known circadian genes, including *Per2 and Rgs16* (M* = 3.11), as well as others, such as *Ak5* and *Erc2,* which are significantly depleted in the SCN in the current study. In a similar way to that of the IEGs, this may indicate the potential for activation from a low resting level of transcription and does not preclude these genes playing important roles in the SCN.

Investigating uncharacterized transcripts

Another consideration for an ongoing meta-analysis to identify tissue-enriched genes is that it provides a strong statistical backing for investigating those transcripts that have received little or no previous attention. Researchers tend to focus on genes and proteins where some research has been recorded previously, and understudied transcripts and genes represent large gaps in knowledge (82,83). Certainly, in the case of these functionally enigmatic genes in the brain, the gaps in knowledge are an effect of when the genes were discovered, rather than any additional complexity or novelty in the genomic or protein sequences (84). As further data are added to such meta-analyses, increasing certainty about expression should lead to study of such transcripts. Now that deposition of transcriptomic data has become part of the established scientific publication process it is likely that meta-analyses like the one described here will continue to grow in power and utility. Furthermore, it seems that this may also be a valuable way to make use of the many well-executed but underpowered microarray and RNA-Seq experiments that may exist in laboratories without ever having reached publication.

Technical considerations

It is important to clarify that the results are heavily-dependent on the way the question has been approached, in that we have compared SCN to the rest of the brain and have prioritized stable transcripts across all studies. Transcripts that show high levels of expression in the SCN and also in other regions of the brain may not be detected as enriched. The same is true of transcripts that cycle rhythmically, even though they may have a peak of expression that would see substantial enrichment in the SCN. With any microarray experiment, it is important to consider the potential sources of error. Aside from the risk of false positives within the significant results the most likely reason for incorrect reporting of SCN-enrichment is likely to be contamination of the sample by surrounding hypothalamic nuclei. The top 20 list of enriched genes contains a number of genes where the hypothalamic pattern of expression is not restricted solely to the SCN (as shown in the data from GENSAT and the Allen Brain Atlas). This may be a reflection of the number of experiments that made use of tissue-punches to isolate the SCN. This technique prioritizes faster isolation of RNA with lower risk of degradation, whereas the use of laser-capture microscopy favours accuracy, but risks bias introduced by uneven sampling across the whole SCN. SCN tissue collected by both methods has been utilized in the meta-analysis.

The variation in the data we have used in the current meta-analysis consists of variation within each of the studies we have defined (SCN and WB data from the same transcriptomic platform) and variance between the studies (τ^2). Whereas the variance in individual studies decreases with the number of samples (array files or RNA-Seq samples), τ^2 will only be decreased with an increasing number of studies. In this respect, limiting the design to one large study per array/transcriptomics platform increases the effects of the between-studies variation. It may be that the costs of such an approach are particularly high when looking at the SCN, where the size and temporal variation in transcription will likely increase the biologically relevant variation between studies. Recent discussions have suggested that, where τ^2 is large, a Fixed Effects Model could be used for the meta-analysis, or more complex Bayesian models could be explored (85). However, with many rhythmic genes present in the data, the large variation between studies may be a sign these genes are both rhythmic and enriched overall. Furthermore, the intent of the current work was to look at enrichment without the assumptions inherent with complex tools or statistical models. I^2 values in the data table (see Supplementary Files 2 and 3) indicate the amount of variation that is not explained by variation within studies (heterogeneity or inconsistency in the meta-analysis (86)) and these may be a better guide to the biological variation of a given transcript (including circadian rhythmicity).

As previously discussed, factors such as the sex, age and strain of the mice used might be expected to change to outcome of the study. Certainly changes in SCN gene expression have also been shown to occur with age (87). For example, SIRT1 levels decline with age, which has been shown to be related to the reduced stability of the molecular clock (88). Sex differences may also alter the function of the SCN (89) and underlying transcription, as could the background strain of the mice used. However, the majority of the samples used for both SCN and WB are from young male mice (up to 16 weeks of age) and no consistent disparity between SCN and WB samples exists in all the studies. Furthermore, accepting these sources of variation into an inclusive meta-analysis is likely to highlight those genes that are consistently enriched in the SCN, regardless of other variables. As more SCN transcriptomic data become available, it may be possible to conduct meta-analyses which further partition data to allow both rhythmic and SCN-enriched genes to be

identified. Genes which show strain, age or sex-related differences in the SCN could be investigated in a similar manner. However, the currently available data is insufficient to allow these factors to be reliably addressed. As such, a comparison of existing data resources to identify rhythmic genes (e.g. CircaDB, Pizarro *et al*. (90)) combined with the SCN enrichment data described here is perhaps the best approach to identify candidate genes by both their circadian expression and SCN enrichment.'

CONCLUSION

Meta-analysis of transcriptomic data provides a powerful approach to identify enriched and depleted transcripts in specific brain regions. Here we apply this approach to the SCN, the master circadian pacemaker of the mammalian hypothalamus. Although many tissues display rhythmic transcription, those transcripts that are consistently enriched in the SCN are likely to have important functions in the generation and maintenance of circadian rhythms. The current analysis has identified both transcripts previously reported as highly expressed in this brain region and genes known to play key roles in SCN physiology, although no single transcript is revealed with high levels of expression restricted solely to the SCN. Moreover, this study identifies a range of transcripts that have, to date, received little attention and have no known SCN function. These provide *bona fide* candidates for future studies to further understand the molecular basis of co-ordinated biological timing.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would also like to thank Sheena Lee, Thomas Vogels, Aarti Jagannath and Michelle Simon for useful discussions during this study.

FUNDING

This work was supported by The Wellcome Trust [098461/Z/12/Z to S.N.P., R.G.F.]. Funding for open access charge: Wellcome Trust.
Conflict of interest statement. None declared.

REFERENCES

1. Moore,R.Y. and Eichler,V.B. (1972) Loss of a circadian adrenal corticosterone rhythm following suprachiasmatic lesions in the rat. *Brain Res.*, **42**, 201–206.
2. Ralph,M.R., Foster,R.G., Davis,F.C. and Menaker,M. (1990) Transplanted suprachiasmatic nucleus determines circadian period. *Science*, **247**, 975–978.
3. Reppert,S.M. and Weaver,D.R. (2002) Coordination of circadian timing in mammals. *Nature*, **418**, 935–941.
4. Colwell,C.S. (2011) Linking neural activity and molecular oscillations in the SCN. *Nat. Rev. Neurosci.*, **12**, 553–569.
5. Herzog,E.D., Hermanstyne,T., Smyllie,N.J. and Hastings,M.H. (2017) Regulating the suprachiasmatic nucleus (SCN) circadian clockwork: interplay between cell-autonomous and circuit-level mechanisms. *Cold Spring Harb. Perspect. Biol.*, **9**, a027706.
6. Steeves,T.D., King,D.P., Zhao,Y., Sangoram,A.M., Du,F., Bowcock,A.M., Moore,R.Y. and Takahashi,J.S. (1999) Molecular cloning and characterization of the human CLOCK gene: expression in the suprachiasmatic nuclei. *Genomics*, **57**, 189–200.
7. Godinho,S.I., Maywood,E.S., Shaw,L., Tucci,V., Barnard,A.R., Busino,L., Pagano,M., Kendall,R., Quwailid,M.M., Romero,M.R. *et al.* (2007) The after-hours mutant reveals a role for Fbxl3 in determining mammalian circadian period. *Science*, **316**, 897–900.
8. Yoo,S.H., Mohawk,J.A., Siepka,S.M., Shan,Y., Huh,S.K., Hong,H.K., Kornblum,I., Kumar,V., Koike,N., Xu,M. *et al.* (2013) Competing E3 ubiquitin ligases govern circadian periodicity by degradation of CRY in nucleus and cytoplasm. *Cell*, **152**, 1091–1105.
9. Parsons,M.J., Brancaccio,M., Sethi,S., Maywood,E.S., Satija,R., Edwards,J.K., Jagannath,A., Couch,Y., Finelli,M.J., Smyllie,N.J. *et al.* (2015) The regulatory factor ZFHX3 modifies circadian function in SCN via an AT motif-driven axis. *Cell*, **162**, 607–621.
10. Tei,H., Okamura,H., Shigeyoshi,Y., Fukuhara,C., Ozawa,R., Hirose,M. and Sakaki,Y. (1997) Circadian oscillation of a mammalian homologue of the Drosophila period gene. *Nature*, **389**, 512–516.
11. Shearman,L.P., Zylka,M.J., Weaver,D.R., Kolakowski,L.F. Jr and Reppert,S.M. (1997) Two period homologs: circadian expression and photic regulation in the suprachiasmatic nuclei. *Neuron*, **19**, 1261–1269.
12. Takumi,T., Taguchi,K., Miyake,S., Sakakida,Y., Takashima,N., Matsubara,C., Maebayashi,Y., Okumura,K., Takekida,S., Yamamoto,S. *et al.* (1998) A light-independent oscillatory gene mPer3 in mouse SCN and OVLT. *EMBO J.*, **17**, 4753–4759.
13. Griffin,E.A. Jr, Staknis,D. and Weitz,C.J. (1999) Light-independent role of CRY1 and CRY2 in the mammalian circadian clock. *Science*, **286**, 768–771.
14. van der Horst,G.T., Muijtjens,M., Kobayashi,K., Takano,R., Kanno,S., Takao,M., de Wit,J., Verkerk,A., Eker,A.P., van Leenen,D. *et al.* (1999) Mammalian Cry1 and Cry2 are essential for maintenance of circadian rhythms. *Nature*, **398**, 627–630.
15. Gekakis,N., Staknis,D., Nguyen,H.B., Davis,F.C., Wilsbacher,L.D., King,D.P., Takahashi,J.S. and Weitz,C.J. (1998) Role of the CLOCK protein in the mammalian circadian mechanism. *Science*, **280**, 1564–1569.
16. Duffield,G.E. (2003) DNA microarray analyses of circadian timing: the genomic basis of biological time. *J. Neuroendocrinol.*, **15**, 991–1002.
17. Panda,S., Antoch,M., Miller,B., Su,A., Schook,A., Straume,M., Schultz,P., Kay,S., Takahashi,J. and Hogenesch,J. (2002) Coordinated transcription of key pathways in the mouse by the circadian clock. *Cell*, **109**, 307–320.
18. Ueda,H.R., Chen,W., Adachi,A., Wakamatsu,H., Hayashi,S., Takasugi,T., Nagano,M., Nakahama,K., Suzuki,Y., Sugano,S. *et al.* (2002) A transcription factor response element for gene expression during circadian night. *Nature*, **418**, 534–539.
19. Reed,H.E., Meyer-Spasche,A., Cutler,D.J., Coen,C.W. and Piggins,H.D. (2001) Vasoactive intestinal polypeptide (VIP) phase-shifts the rat suprachiasmatic nucleus clock in vitro. *Eur. J. Neurosci.*, **13**, 839–843.
20. Harmar,A.J., Marston,H.M., Shen,S., Spratt,C., West,K.M., Sheward,W.J., Morrison,C.F., Dorin,J.R., Piggins,H.D., Reubi,J.C. *et al.* (2002) The VPAC(2) receptor is essential for circadian function in the mouse suprachiasmatic nuclei. *Cell*, **109**, 497–508.
21. Freeman,G.M. Jr, Krock,R.M., Aton,S.J., Thaben,P. and Herzog,E.D. (2013) GABA networks destabilize genetic oscillations in the circadian pacemaker. *Neuron*, **78**, 799–806.
22. Kasukawa,T., Masumoto,K.-h., Nikaido,I., Nagano,M., Uno,K., Tsujino,K., Hanashima,C., Shigeyoshi,Y. and Ueda,H. (2011) Quantitative expression profile of distinct functional regions in the adult mouse brain. *PLoS One*, **6**, e23228.
23. VanDunk,C., Hunter,L. and Gray,P. (2011) Development, maturation, and necessity of transcription factors in the mouse suprachiasmatic nucleus. *J. Neurosci.*, **31**, 6457–6467.
24. Husse,J., Zhou,X., Shostak,A., Oster,H. and Eichele,G. (2011) Synaptotagmin10-Cre, a driver to disrupt clock genes in the SCN. *J. Biol. Rhythms*, **26**, 379–389.
25. Jones,J.R., Tackenberg,M.C. and McMahon,D.G. (2015) Manipulating circadian clock neuron firing rate resets molecular circadian rhythms and behavior. *Nat. Neurosci.*, **18**, 373–375.

26. Nadon,R. and Shoemaker,J. (2002) Statistical issues with microarrays: processing and analysis. *Trends Genet.*, **18**, 265–271.
27. Chalmers,I. (1993) The cochrane collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. *Ann. N. Y. Acad. Sci.*, **703**, 156–163.
28. Edgar,R., Domrachev,M. and Lash,A. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
29. Brazma,A., Parkinson,H., Sarkans,U., Shojatalab,M., Vilo,J., Abeygunawardena,N., Holloway,E., Kapushesky,M., Kemmeren,P., Lara,G. *et al.* (2003) ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, **31**, 68–71.
30. Rhodes,D., Yu,J., Shanker,K., Deshpande,N., Varambally,R., Ghosh,D., Barrette,T., Pandey,A. and Chinnaiyan,A. (2004) Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 9309–9314.
31. Breitling,R., Armengaud,P., Amtmann,A. and Herzyk,P. (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.*, **573**, 83–92.
32. Rhodes,D., Barrette,T., Rubin,M., Ghosh,D. and Chinnaiyan,A. (2002) Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res.*, **62**, 4427–4433.
33. Choi,J.K., Yu,U., Kim,S. and Yoo,O.J. (2003) Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, **19**, i84–i90.
34. Ramasamy,A., Mondry,A., Holmes,C.C. and Altman,D.G. (2008) Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med.*, **5**, e184.
35. Borenstein,M. (2009) *Introduction to Meta-Analysis*. John Wiley & Sons, Chichester.
36. Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
37. Robinson,J.T., Thorvaldsdottir,H., Winckler,W., Guttman,M., Lander,E.S., Getz,G. and Mesirov,J.P. (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.
38. Emig,D., Salomonis,N., Baumbach,J., Lengauer,T., Conklin,B.R. and Albrecht,M. (2010) AltAnalyze and DomainGraph: analyzing and visualizing exon expression data. *Nucleic Acids Res.*, **38**, W755–W762.
39. Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 9440–9445.
40. Team,R.C. (2015). *R Foundation for Statistical Computing*. R Foundation for Statistical Computing, Vienna.
41. Storey,J., Bass,J.D., Swcf,A.J., Dabney,A. and Robinson,D. (2015) qvalue: Q-value estimation for false discovery rate control. R package version 2.0.0. http://github.com/jdstorey/qvalue.
42. Lein,E., Hawrylycz,M., Ao,N., Ayres,M., Bensinger,A., Bernard,A., Boe,A., Boguski,M., Brockway,K., Byrnes,E. *et al.* (2007) Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, **445**, 168–176.
43. Gong,S., Zheng,C., Doughty,M.L., Losos,K., Didkovsky,N., Schambra,U.B., Nowak,N.J., Joyner,A., Leblanc,G. and Hatten,M.E. (2003) A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature*, **425**, 917–925.
44. Shannon,P., Markiel,A., Ozier,O., Baliga,N., Wang,J., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
45. Szklarczyk,D., Franceschini,A., Wyder,S., Forslund,K., Heller,D., Huerta-Cepas,J., Simonovic,M., Roth,A., Santos,A., Tsafou,K.P. *et al.* (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, **43**, D447–D452.
46. Maere,S., Heymans,K. and Kuiper,M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.
47. Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.
48. Merico,D., Isserlin,R., Stueker,O., Emili,A. and Bader,G.D. (2010) Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One*, **5**, e13984.
49. Yi,Z., Yokota,H., Torii,S., Aoki,T., Hosaka,M., Zhao,S., Takata,K., Takeuchi,T. and Izumi,T. (2002) The Rab27a/granuphilin complex regulates the exocytosis of insulin-containing dense-core granules. *Mol. Cell. Biol.*, **22**, 1858–1867.
50. Schou,K.B., Pedersen,L.B. and Christensen,S.T. (2015) Ins and outs of GPCR signaling in primary cilia. *EMBO Rep.*, **16**, 1099–1113.
51. Ramkisoensing,A. and Meijer,J.H. (2015) Synchronization of biological clock neurons by light and peripheral feedback systems promotes circadian rhythms and health. *Front. Neurol.*, **6**, 128.
52. Mieda,M., Ono,D., Hasegawa,E., Okamoto,H., Honma,K., Honma,S. and Sakurai,T. (2015) Cellular clocks in AVP neurons of the SCN are critical for interneuronal coupling regulating circadian behavior rhythm. *Neuron*, **85**, 1103–1116.
53. Mori,K., Miyazato,M., Ida,T., Murakami,N., Serino,R., Ueta,Y., Kojima,M. and Kangawa,K. (2005) Identification of neuromedin S and its possible role in the mammalian circadian oscillator system. *EMBO J.*, **24**, 325–335.
54. Cheng,M.Y., Bullock,C.M., Li,C., Lee,A.G., Bermak,J.C., Belluzzi,J., Weaver,D.R., Leslie,F.M. and Zhou,Q.Y. (2002) Prokineticin 2 transmits the behavioural circadian rhythm of the suprachiasmatic nucleus. *Nature*, **417**, 405–410.
55. Hundahl,C.A., Fahrenkrug,J., Hay-Schmidt,A., Georg,B., Faltoft,B. and Hannibal,J. (2012) Circadian behaviour in neuroglobin deficient mice. *PLoS One*, **7**, e34462.
56. Kozlov,S.V., Bogenpohl,J.W., Howell,M.P., Wevrick,R., Panda,S., Hogenesch,J.B., Muglia,L.J., Van Gelder,R.N., Herzog,E.D. and Stewart,C.L. (2007) The imprinted gene Magel2 regulates normal circadian output. *Nat. Genet.*, **39**, 1266–1272.
57. Devos,J., Weselake,S.V. and Wevrick,R. (2011) Magel2, a Prader-Willi syndrome candidate gene, modulates the activities of circadian rhythm proteins in cultured cells. *J. Circadian Rhythms*, **9**, 12.
58. Jagannath,A., Butler,R., Godinho,S.I., Couch,Y., Brown,L.A., Vasudevan,S.R., Flanagan,K.C., Anthony,D., Churchill,G.C., Wood,M.J. *et al.* (2013) The CRTC1-SIK1 pathway regulates entrainment of the circadian clock. *Cell*, **154**, 1100–1111.
59. Porterfield,V., Piontkivska,H. and Mintz,E. (2007) Identification of novel light-induced genes in the suprachiasmatic nucleus. *BMC Neurosci.*, **8**, 98.
60. Araki,R., Nakahara,M., Fukumura,R., Takahashi,H., Mori,K., Umeda,N., Sujino,M., Inouye,S.T. and Abe,M. (2006) Identification of genes that express in response to light exposure and express rhythmically in a circadian manner in the mouse suprachiasmatic nucleus. *Brain Res.*, **1098**, 9–18.
61. DeWoskin,D., Myung,J., Belle,M.D., Piggins,H.D., Takumi,T. and Forger,D.B. (2015) Distinct roles for GABA across multiple timescales in mammalian circadian timekeeping. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E3911–E3919.
62. Myung,J., Hong,S., DeWoskin,D., De Schutter,E., Forger,D.B. and Takumi,T. (2015) GABA-mediated repulsive coupling between circadian clock neurons in the SCN encodes seasonal time. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E3920–E3929.
63. Fan,J., Zeng,H., Olson,D.P., Huber,K.M., Gibson,J.R. and Takahashi,J.S. (2015) Vasoactive intestinal polypeptide (VIP)-expressing neurons in the suprachiasmatic nucleus provide sparse GABAergic outputs to local neurons with circadian regulation occurring distal to the opening of postsynaptic GABAA ionotropic receptors. *J. Neurosci.*, **35**, 1905–1920.
64. Abrahamson,E.E. and Moore,R.Y. (2001) Suprachiasmatic nucleus in the mouse: retinal innervation, intrinsic organization and efferent projections. *Brain Res.*, **916**, 172–191.
65. Deery,M.J., Maywood,E.S., Chesham,J.E., Sladek,M., Karp,N.A., Green,E.W., Charles,P.D., Reddy,A.B., Kyriacou,C.P., Lilley,K.S. *et al.* (2009) Proteomic analysis reveals the role of synaptic vesicle cycling in sustaining the suprachiasmatic circadian clock. *Curr. Biol.*, **19**, 2031–2036.
66. Kim,S.M., Power,A., Brown,T.M., Constance,C.M., Coon,S.L., Nishimura,T., Hirai,H., Cai,T., Eisner,C., Weaver,D.R. *et al.* (2009) Deletion of the secretory vesicle proteins IA-2 and IA-2beta disrupts circadian rhythms of cardiovascular and physical activity. *FASEB J.*, **23**, 3226–3232.

67. Lyakhova,T.A. and Knight,J.D. (2014) The C2 domains of granuphilin are high-affinity sensors for plasma membrane lipids. *Chem. Phys. Lipids*, **182**, 29–37.
68. Ostrowski,M., Carmo,N.B., Krumeich,S., Fanget,I., Raposo,G., Savina,A., Moita,C.F., Schauer,K., Hume,A.N., Freitas,R.P. *et al.* (2010) Rab27a and Rab27b control different steps of the exosome secretion pathway. *Nat. Cell Biol.*, **12**, 19–30.
69. Xu,X., Coats,J.K., Yang,C.F., Wang,A., Ahmed,O.M., Alvarado,M., Izumi,T. and Shah,N.M. (2012) Modular genetic control of sexually dimorphic behaviors. *Cell*, **148**, 596–607.
70. Evans,J.A., Leise,T.L., Castanon-Cervantes,O. and Davidson,A.J. (2013) Dynamic interactions mediated by nonredundant signaling mechanisms couple circadian clock neurons. *Neuron*, **80**, 973–983.
71. Twelvetrees,A.E., Yuen,E.Y., Arancibia-Carcamo,I.L., MacAskill,A.F., Rostaing,P., Lumb,M.J., Humbert,S., Triller,A., Saudou,F., Yan,Z. *et al.* (2010) Delivery of GABAARs to synapses is mediated by HAP1-KIF5 and disrupted by mutant huntingtin. *Neuron*, **65**, 53–65.
72. Walling,H.W., Baldassare,J.J. and Westfall,T.C. (1998) Molecular aspects of Huntington's disease. *J. Neurosci. Res.*, **54**, 301–308.
73. Morton,A.J., Wood,N.I., Hastings,M.H., Hurelbrink,C., Barker,R.A. and Maywood,E.S. (2005) Disintegration of the sleep-wake cycle and circadian timing in Huntington's disease. *J. Neurosci.*, **25**, 157–163.
74. Shi,S.Q., Bichell,T.J., Ihrie,R.A. and Johnson,C.H. (2015) Ube3a imprinting impairs circadian robustness in Angelman syndrome models. *Curr. Biol.*, **25**, 537–545.
75. Ehlen,J.C., Jones,K.A., Pinckney,L., Gray,C.L., Burette,S., Weinberg,R.J., Evans,J.A., Brager,A.J., Zylka,M.J., Paul,K.N. *et al.* (2015) Maternal Ube3a loss disrupts sleep homeostasis but leaves circadian rhythmicity largely intact. *J. Neurosci.*, **35**, 13587–13598.
76. Lassi,G., Ball,S.T., Maggi,S., Colonna,G., Nieus,T., Cero,C., Bartolomucci,A., Peters,J. and Tucci,V. (2012) Loss of Gnas imprinting differentially affects REM/NREM sleep and cognition in mice. *PLoS Genet.*, **8**, e1002706.
77. Gregg,C., Zhang,J., Weissbourd,B., Luo,S., Schroth,G.P., Haig,D. and Dulac,C. (2010) High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science*, **329**, 643–648.
78. Bonthuis,P.J., Huang,W.C., Stacher Horndli,C.N., Ferris,E., Cheng,T. and Gregg,C. (2015) Noncanonical genomic imprinting effects in offspring. *Cell Rep.*, **12**, 979–991.
79. Perez,J.D., Rubinstein,N.D., Fernandez,D.E., Santoro,S.W., Needleman,L.A., Ho-Shing,O., Choi,J.J., Zirlinger,M., Chen,S.K., Liu,J.S. *et al.* (2015) Quantitative and functional interrogation of parent-of-origin allelic expression biases in the brain. *Elife*, **4**, e07860.
80. Lane,J.M., Vlasac,I., Anderson,S.G., Kyle,S.D., Dixon,W.G., Bechtold,D.A., Gill,S., Little,M.A., Luik,A., Loudon,A. *et al.* (2016) Genome-wide association analysis identifies novel loci for chronotype in 100,420 individuals from the UK Biobank. *Nat. Commun.*, **7**, 10889.
81. Hu,Y., Shmygelska,A., Tran,D., Eriksson,N., Tung,J.Y. and Hinds,D.A. (2016) GWAS of 89,283 individuals identifies genetic variants associated with self-reporting of being a morning person. *Nat. Commun.*, **7**, 10448.
82. Edwards,A.M., Isserlin,R., Bader,G.D., Frye,S.V., Willson,T.M. and Yu,F.H. (2011) Too many roads not taken. *Nature*, **470**, 163–165.
83. Su,A.I. and Hogenesch,J.B. (2007) Power-law-like distributions in biomedical publications and research funding. *Genome Biol.*, **8**, 404–405.
84. Pandey,A.K., Lu,L., Wang,X., Homayouni,R. and Williams,R.W. (2014) Functionally enigmatic genes: a case study of the brain ignorome. *PLoS One*, **9**, e88889.
85. Borenstein,M., Hedges,L.V., Higgins,J.P. and Rothstein,H.R. (2010) A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res. Synth. Methods*, **1**, 97–111.
86. Higgins,J.P., Thompson,S.G., Deeks,J.J. and Altman,D.G. (2003) Measuring inconsistency in meta-analyses. *BMJ*, **327**, 557–560.
87. Banks,G., Nolan,P.M. and Peirson,S.N. (2016) Reciprocal interactions between circadian clocks and aging. *Mamm. Genome*, **27**, 332–340.
88. Chang,H.C. and Guarente,L. (2013) SIRT1 mediates central circadian control in the SCN by a mechanism that decays with aging. *Cell*, **153**, 1448–1460.
89. Kuljis,D.A., Loh,D.H., Truong,D., Vosko,A.M., Ong,M.L., McClusky,R., Arnold,A.P. and Colwell,C.S. (2013) Gonadal- and sex-chromosome-dependent sex differences in the circadian system. *Endocrinology*, **154**, 1501–1512.
90. Pizarro,A., Hayer,K., Lahens,N.F. and Hogenesch,J.B. (2013) CircaDB: a database of mammalian circadian gene expression profiles. *Nucleic Acids Res.*, **41**, D1009–D1013.
91. Thorrez,L., Laudadio,I., Van Deun,K., Quintens,R., Hendrickx,N., Granvik,M., Lemaire,K., Schraenen,A., Van Lommel,L., Lehnert,S. *et al.* (2011) Tissue-specific disallowance of housekeeping genes: the other face of cell differentiation. *Genome Res.* **21**, 95–105.
92. Bhave,S., Hoffman,P., Lassen,N., Vasiliou,V., Saba,L., Deitrich,R. and Tabakoff,B. (2006) Gene array profiles of alcohol and aldehyde metabolizing enzymes in brains of C57BL/6 and DBA/2 mice. *Alcohol. Clin. Exp. Res.*, **30**, 1659–1669.
93. Doi,M., Ishida,A., Miyake,A., Sato,M., Komatsu,R., Yamazaki,F., Kimura,I., Tsuchiya,S., Kori,H. and Seo,K. (2011) Circadian regulation of intracellular G-protein signalling mediates intercellular synchrony and rhythmicity in the suprachiasmatic nucleus. *Nat. Commun.*, **2**, 327.
94. Kedmi,M. and Orr-Urtreger,A. (2011) The effects of aging vs. α7 nAChR subunit deficiency on the mouse brain transcriptome: aging beats the deficiency. *Age (Dordr)*, **33**, 1–13.
95. Kleiber,M., Laufer,B., Wright,E., Diehl,E. and Singh,S. (2012) Long-term alterations to the brain transcriptome in a maternal voluntary consumption model of fetal alcohol spectrum disorders. *Brain Res.*, **1458**, 18–33.
96. Laderas,T., Walter,N., Mooney,M., Vartanian,K., Darakjian,P., Buck,K., Harrington,C., Belknap,J., Hitzemann,R. and McWeeney,S. (2011) Computational detection of alternative exon usage. *Front. Neurosci.*, **5**, 69.
97. Lovegrove,F., Gharib,S., Patel,S., Hawkes,C., Kain,K. and Liles,W. (2007) Expression microarray analysis implicates apoptosis and interferon-responsive mechanisms in susceptibility to experimental cerebral malaria. *Am. J. Pathol.*, **171**, 1894–1903.
98. Oliver,P., Sobczyk,M., Maywood,E., Edwards,B., Lee,S., Livieratos,A., Oster,H., Butler,R., Godinho,S., Wulff,K. *et al.* (2012) Disrupted circadian rhythms in a mouse model of schizophrenia. *Curr. Biol.*, **22**, 314–319.
99. Azzi,A., Dallmann,R., Casserly,A., Rehrauer,H., Patrignani,A., Maier,B., Kramer,A. and Brown,S.A. (2014) Circadian behavior is light-reprogrammed by plastic DNA methylation. *Nat. Neurosci.*, **17**, 377–382.
100. Fushan,A.A., Turanov,A.A., Lee,S.G., Kim,E.B., Lobanov,A.V., Yim,S.H., Buffenstein,R., Lee,S.R., Chang,K.T., Rhee,H. *et al.* (2015) Gene expression defines natural changes in mammalian lifespan. *Aging Cell*, **14**, 352–365.
101. Brawand,D., Soumillon,M., Necsulea,A., Julien,P., Csardi,G., Harrigan,P., Weier,M., Liechti,A., Aximu-Petri,A., Kircher,M. *et al.* (2011) The evolution of gene expression levels in mammalian organs. *Nature*, **478**, 343–348.
102. Yalcin,B., Wong,K., Agam,A., Goodson,M., Keane,T.M., Gan,X., Nellaker,C., Goodstadt,L., Nicod,J., Bhomra,A. *et al.* (2011) Sequence-based characterization of structural variation in the mouse genome. *Nature*, **477**, 326–329.
103. Doi,M., Ishida,A., Miyake,A., Sato,M., Komatsu,R., Yamazaki,F., Kimura,I., Tsuchiya,S., Kori,H., Seo,K. *et al.* (2011) Circadian regulation of intracellular G-protein signalling mediates intercellular synchrony and rhythmicity in the suprachiasmatic nucleus. *Nat. Commun.*, **2**, 327.

OPEN BIOLOGY

royalsocietypublishing.org/journal/rsob

Review

Cite this article: Jahangiri L, Tsaprouni L, Trigg RM, Williams JA, Gkoutos GV, Turner SD, Pereira J. 2020 Core regulatory circuitries in defining cancer cell identity across the malignant spectrum. *Open Biol.* **10**: 200121. http://dx.doi.org/10.1098/rsob.200121

Received: 7 May 2020
Accepted: 18 June 2020

Subject Area:
genomics

Keywords:
core regulatory circuitry, liquid and solid cancers, super-enhancers, cell identity

Author for correspondence:
Leila Jahangiri
e-mail: leila.jahangiri@bcu.ac.uk

THE ROYAL SOCIETY PUBLISHING

Core regulatory circuitries in defining cancer cell identity across the malignant spectrum

Leila Jahangiri[1,2], Loukia Tsaprouni[1], Ricky M. Trigg[2,3], John A. Williams[4,5,6], Georgios V. Gkoutos[4,5,7,8,9,10], Suzanne D. Turner[2] and Joao Pereira[11]

[1]Department of Life Sciences, Birmingham City University, Birmingham, UK
[2]Division of Cellular and Molecular Pathology, Addenbrooke's Hospital, University of Cambridge, Cambridge, UK
[3]Department of Functional Genomics, GlaxoSmithKline, Stevenage, UK
[4]Institute of Translational Medicine, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK
[5]Institute of Cancer and Genomic Sciences, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK
[6]Mammalian Genetics Unit, Medical Research Council Harwell Institute, Oxfordshire, UK
[7]MRC Health Data Research, UK
[8]NIHR Experimental Cancer Medicine Centre, Birmingham, UK
[9]NIHR Surgical Reconstruction and Microbiology Research Centre, Birmingham, UK
[10]NIHR Biomedical Research Centre, Birmingham, UK
[11]Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Charlestown, USA

LJ, 0000-0003-0235-8447; JAW, 0000-0002-0357-5454; SDT, 0000-0002-8439-4507

Gene expression programmes driving cell identity are established by tightly regulated transcription factors that auto- and cross-regulate in a feed-forward manner, forming core regulatory circuitries (CRCs). CRC transcription factors create and engage super-enhancers by recruiting acetylation writers depositing permissive H3K27ac chromatin marks. These super-enhancers are largely associated with BET proteins, including BRD4, that influence higher-order chromatin structure. The orchestration of these events triggers accessibility of RNA polymerase machinery and the imposition of lineage-specific gene expression. In cancers, CRCs drive cell identity by superimposing developmental programmes on a background of genetic alterations. Further, the establishment and maintenance of oncogenic states are reliant on CRCs that drive factors involved in tumour development. Hence, the molecular dissection of CRC components driving cell identity and cancer state can contribute to elucidating mechanisms of diversion from pre-determined developmental programmes and highlight cancer dependencies. These insights can provide valuable opportunities for identifying and re-purposing drug targets. In this article, we review the current understanding of CRCs across solid and liquid malignancies and avenues of investigation for drug development efforts. We also review techniques used to understand CRCs and elaborate the indication of discussed CRC transcription factors in the wider context of cancer CRC models.

1. Introduction

Programmes involved in the control of gene expression governing cell state, cell state transitions and cellular identity across cell types or lineages have not been comprehensively defined. However, multiple efforts encompassing a myriad of differentiation models have shed light on the mechanisms regulating these developmental programmes [1–5]. These programmes are controlled by a small set of tightly regulated transcription factors (TFs) and/or *de novo* fusion chimeric TFs, forming core regulatory circuitries (CRCs). These CRCs control lineage-specific flow of information for gene expression [6–8]. Mechanistically,

these core regulatory TFs (CR TFs) can control the placement of acetylation deposits around an array of CR TF binding motifs by recruiting acetylation writers, readers and erasers, thereby creating super-enhancers (SEs) [9]. SEs are broad, spatially co-localized enhancer regions that recruit dense transcriptional machinery. SEs are disproportionately larger than most enhancer domains and contain close to 40% of enhancer-associated factors (including epigenetic machinery), while comprising only 3–5% of enhancer regions [10]. CR TFs drive cell identity by binding to SEs associated with lineage identity imposing genes, often oncogenes [6,8,10–12]. CR TFs self-regulate and, they inwardly bind to their own regulatory regions and mutually regulate within the CRC, forming a cross-regulated feed-forward loop [6]. Research efforts to date have focused on understanding components of CRCs and their roles in multiple cell types, including embryonic stem cells (ESCs), induced pluripotent stem cells (iPSCs) and multiple cancer cell types [13–15]. In ESCs, CRC TFs including OCT4, SOX2 and NANOG regulate themselves and each other [10,14]. These CRC TFs dominate the transcriptional programmes governing stem cell self-renewal, pluripotency and cell fate [10,14]. Expression of this network of CRC TFs, with the addition of the proto-oncogene C-MYC, was sufficient to reprogramme somatic cells into iPSCs [16]. Similar efforts in cancers have brought into focus tumour dependencies and regulatory diversity and, in some cases, addiction to regulatory circuitries [15]. Further, SEs, as components of CRCs, are linked to regions of somatic genetic alterations such as focal amplifications in cancers and disease linked-SNPs [17,18]. SEs can also reinforce the expression of factors indicated in tumour development and progression [11].

An important step in understanding the role of CRCs in cancers is the systematic reconstruction of CRCs both in development and cancer. The reconstruction of CRCs for a cell type requires SE maps (usually indicated by high levels of a H3K27ac histone signature), core TF binding data, their putative binding sites in the SE regions and their extended, genome-wide, regulatory network [6,19]. To that end, Saint-André and colleagues reconstructed and predicted CRC models using a CRC mapper programme for 75 human cell and tissue types [6]. Huang and colleagues developed a dbCoRC database which, in addition to archiving CRC information, interactively reconstructs CRCs for over 230 human and mouse cell lines or primary tissue, inclusive of 79 cancer cells and tissues [19]. This database provides cell-type specific information about SEs, CRC models, putative binding sites for TFs identified in target gene SEs, and TF expression patterns [19]. Other resources such as dbSUPER also provide a comprehensive map of SEs identified in more than 100 cell types, which may be used to complement CRC model data [20]. The next step beyond CRC reconstruction in cancers is understanding the cellular and molecular mechanisms of divergence of constitutive developmental programmes in a background of genetic aberrations [6]. The inference of the underlying transcriptional networks that regulate physiological and pathological states is likely to inform these mechanisms of diversion and enhance our understanding of both physiology and disease. Put together, it is reasonable to propose that understanding the role of CRCs in cancers will facilitate the dissection of identity-conferring programmes and lead to a better understanding of their deregulation in cancers, potentially informing drug development and re-purposing strategies [15,21,22]. In this article, we review the present knowledge of CRCs across a multitude of solid and liquid cancers, and the current evidence for leveraging this information for therapeutic gain. We then attempt to elaborate the indication of discussed CRC TFs, in a wider range of cancer cells and tissues using the dbCoRC database. Finally, we describe current methodologies used to understand CRCs.

2. CRCs in a multitude of solid and liquid cancer types

In this section, we address the role of CRCs in controlling the flow of information that governs identity-conferring programmes in a multitude of solid and liquid cancer types (figure 1).

2.1. Neuroblastoma

Neuroblastoma (NB) is a solid malignancy derived from multipotent neural crest cells (NCCs) and contributes to 15% of cancer-related mortality in children [23]. Recent studies have defined the presence of two interconvertible types of NBs regulated by CRCs; committed adrenergic (ADRN) and neural crest migratory (or mesenchymal; MES) [12,24]. Though both cell populations are oncogenic [24], the latter type displays greater therapeutic resistance and encompasses the majority of relapsed tumours [25].

The Notch signalling pathway is the driver of motile MES identity, consistent with a mesenchymal phenotype. MES CRCs include the NOTCH receptors and cofactors, *NOTCH2* and *MAML2*, respectively, which are associated with SEs and drive an array of NOTCH target genes including *HES1* [24,26,27]. Members of the CRC-regulating MES state, namely, the NOTCH family, NOTCH1, NOTCH2 and NOTCH3, can initiate transdifferentiation to the ADRN state through H3K27ac landscape remodelling [24] and hence control maintenance of the MES state. However, the intracellular domain of NOTCH3 is the strongest inducer of reprogramming towards the MES state. Induction of the NOTCH3 intracellular domain leads to *de novo* establishment of SEs at *NOTCH2* and *MAML2* loci as well as the deposition of H3K27ac at the promoter regions of *JAG1*, *NOTCH1*, *NOTCH3* and *HES1* [24].

The CRC regulating the ADRN subtype in NB comprises PHOX2B, HAND2, TBX2, ISL1, ASCL1 and GATA3, whose effects are amplified by MYCN and LMO1 [25,28–30]. The most recent addition to this circuitry, ASCL1, a bHLH transcription factor implicated in NB cell growth and differentiation arrest, is directly regulated by LMO1, MYCN and other members of the CRC [31]. Similarly, ASCL1 directly regulates the expression of other genes in this CRC, forming an auto-regulatory loop [31]. Other members of this CRC, including GATA3, a biomarker linked to the proliferation of NB cells and self-renewal capacity [32], is downregulated following retinoic acid (RA) treatment, inhibiting tumourigenicity [32,33]. In addition, ISL1 positively regulates cell cycle genes and represses genes associated with differentiation (e.g. RA receptors, *CDKN1A* and *EPAS1*) [34].

The events leading to the oncogenic capacity and specificity of both ADRN and MES NB subtypes during

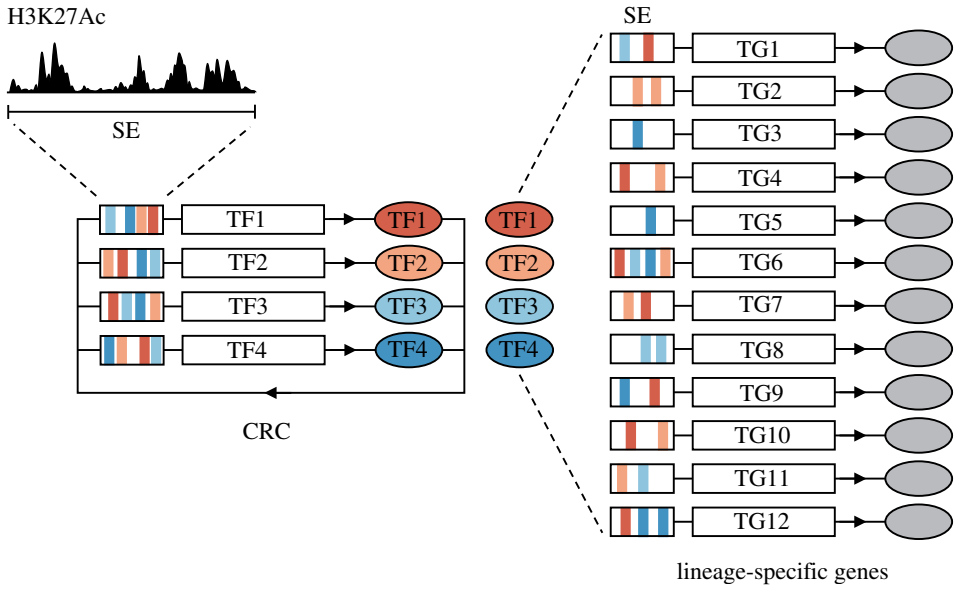


Figure 1. Core regulatory circuitry (CRC) constitutes a network that can confer lineage-specific gene expression. Core regulatory transcription factors (CR TFs) self-regulate and regulate the expression of other CR TFs in a cross-regulated feed-forward loop. Super-enhancers (SEs) contain CR TF binding sites and marked H3K27ac deposits. CR TFs, in turn, bind to the regulatory regions of a network of target genes (TGs) including lineage-specific genes that drive cell identity. TF, transcription factor; TG, target gene; SE, super enhancer. TF binding motifs (for TF1, TF2, TF3 and TF4) are depicted as rectangles in red, pink, light blue and dark blue, respectively.

development are still unknown. However, recent work by Soldatov and colleagues, which profiled gene expression during mouse neural crest development, may provide insights into the timing of NB oncogenesis. Single-cell RNA sequencing identified a novel bipotent cell type, a dual fate progenitor expressing both Phox2b and Prrx1, late in the differentiation cascade of NCCs [35]. As discussed, PHOX2B is expressed in ADRN subtypes while PRRX1 is MES-specific, and its overexpression is sufficient to convert ADRN to MES subtypes [24,25]. The existence of these dual progenitors could indicate they are upstream of the oncogenic event leading to the formation of both MES and ADRN NBs, and that further characterization of the complex SEs regulating cell fate decisions at this stage will be likely to inform NB biology. Table 1 summarizes examples of CRC TFs discussed in this section.

2.2. Glioblastoma

Glioblastoma (GBM) is the most common primary malignant brain tumour in adults and harbours distinct heterogeneous populations of tumour cells [43]. Earlier studies identified CRCs comprising the POU3F2, SOX2 and SALL2, OLIG2 TFs whose activities reprogrammed differentiated GBM cells into induced tumour propagating cells (TPCs). These TPCs have stem-like properties, are capable of tumourigenesis and display unique SE landscapes [43–45]. A target gene of this network is RCOR2, which forms a protein complex with LSD1, a histone methyltransferase. The RCOR2/LSD1 complex replaces OLIG2 in the reprogramming cocktail towards TPC [44]. Notably, most of these genes are involved in the maintenance of neural stem cell (NSC) identity during development. Expression of Pou3f2 (Brn2) was shown to be sufficient to convert astrocytes into neural progenitors in mice, similar to its role in the formation of TPCs [46]; SOX2 and OLIG2 are involved in maintaining the identity and replication potential of neural progenitors [47,48].

In a study conducted on glioblastoma stem cells (GSCs), NOTCH1, SOX2, SALL2, POU3F and OLIG2 blocked differentiation in GSCs, confirming the observations made in GBM by Suvà and colleagues [44,45]. Although the similarities and differences between induced TPCs and GSCs is not clear, it may be possible to propose that cells with self-renewal and tumourigenesis capacity can be identified in GBM or induced from differentiated GBM. Building on these observations, in a more recent study, Riddick and colleagues compare the global gene expression pattern of GSCs and NSCs during *in vitro* differentiation [36]. This group revealed a substantial overlap between the regulatory landscape of GSCs and NSCs. Further, in addition to the identification of important transcriptional regulators of GSC and NSC biology, such as SOX2, OLIG2, DLL, NOTCH and HES1, there were other significant observations. First, GSCs akin to NSCs express SOX2, Nestin and CD133, and demonstrate self-renewal and multi-potency while sharing common yet deregulated developmental pathways with NSCs including AKT, RAS, NOTCH, BMI-1 and WNT [36,49–53]. Second, the binding signature of TFs to differentially expressed genes was used to reconstruct a CRC centred on KLF4, a TF involved in activation of DDL1, NOTCH1 and SOX2 [36]. The overexpression of KLF4 in both GSCs and NSCs blocks differentiation and reduces proliferation [36,54]. In GSCs, KLF4 is regulated by ERG1 and sits downstream of STAT3 in the PI3K pathway [36].

Finally, consistent with potential plasticity of cell identity, glioblastomas can be reprogrammed towards mesenchymal lineages by the synergistic activity of initiators and master regulators, including STAT3 (downstream of PI3K activity) and CEBPB. Ectopic expression of these genes in NSCs reprogrammes these cells towards the mesenchymal lineages, and their expression in tumours is predictive of poor clinical outcomes, consistent with promoting motile phenotypes in these cells [55]. Table 1 summarizes examples of CRC TFs discussed in this section.

Table 1. Summary data of relevant CRC TFs identified in the indicated malignancies. In this table, cancer type and examples of subtype, subgroups, cliques or modules identified have been summarized. Further examples of CRC TF identified in each subtype, group, module or clique have been provided.

cancer	subtypes, subgroups or identified modules	examples of identified CRC TFs
neuroblastoma [24,31]	MES	NOTCH receptors and cofactors including NOTCH2 and MAML2
	ADRN	PHOX2B, HAND2, TBX2, ISL1, ASCL1 and GATA3, MYCN and LMO1
glioblastoma [36]		KLF4, ERG1, Notch pathway and SOX2
rhabdomyosarcoma [9]	Pan-RMS	MYOD1 and MYOG
	FP-RMS	PAX3-FOXO1, MYCN, SOX8, MYOD1 and MYOG
	fusion–negative RMS	PAX7 and AP1 family
	normal muscle-specific (NMS)	Nur77 and MEF2D
renal cell carcinoma [37]		PAX8
liposarcoma [38]	myxoid (MLPS)	FUS-DDIT3
	de-differentiated (DDLPS)	FOSL2, MYC and RUNX1
prostate cancer [39]		AR and ERG
gastrointestinal stromal tumour (GIST) [40]		FOXF1 and ETV1
medullablastoma [21]	group 3	HLX and LHX2
	group 4	LMX1A and LHX2
chronic lymphocytic leukaemia (CLL) [41]	CLL-2 (clique 2)	PAX5, ETV6, TCF3, IRF2, MEF2D, ELF1, KLF13, JUND, FOXP1, IRF1 and IRF8
	CLL-11 (clique 11)	PAX5, ETV6, TCF12, IRF2, RARA, NFATC1, KLF12, JUN, RUNX3 and FLI1
T-cell acute lymphoblastic leukaemia (T-ALL) [42]		TAL1, GATA3 and RUNX1

2.3. Rhabdomyosarcoma

Childhood rhabdomyosarcoma (RMS) is the most common soft tissue sarcoma in paediatric patients [56]. RMS oncogenesis relies on the expression of myogenic TFs [57], generating at least four identified CRCs in RMS tissue and cell lines: (i) a pan-RMS CRC defined by expression of MYOD1 and MYOG; (ii) a fusion-positive RMS (FP-RMS), which includes FOXO1 (SEs regulating *PAX3-FOXO1* or *PAX7-FOXO1*) and MYCN; (iii) a fusion-negative RMS including PAX7 and the AP1 family of TFs; and (iv) a normal muscle-specific CRC with TFs expressing Nur77 and MEF2D [58,59].

The FP-RMS module is formed by a t(2:13)(q35:q14) translocation forming a *PAX3-FOXO1* fusion gene, which functions as a primary oncogenic driver [9]. A consistently high-scoring H3K27ac signal and open chromatin structure was identified in the SE regions of *SOX8* in primary FP-RMS samples. More detailed investigation revealed that PAX3-FOXO1 positively regulates MYOD1, MOYG and SOX8 in a feed-forward mechanism [9].

MYOD1 and MYOG lead a pro-myogenic programme in RMS, while SOX8, a regulator of early neural crest development, displays anti-myogenic functions and opposes the ability of these factors to complete muscle differentiation [60]. Crucially, it is through the binding of PAX3-FOXO1 to SEs of *SOX8* and subsequent activation of SOX8 expression that this fusion protein can exert its anti-differentiation activity on these cells [9]. In conclusion, MYOD1 and MYOG are drivers of the myogenic programme, which is opposed by PAX3-FOXO1 via binding to the SE of *SOX8*.

The transcriptional interaction between SOX8, MYOD1 and MYOG is also interesting. Disruption of either MYOD1 or MYOG results in dramatic transcriptional downregulation of *MYOD1*, *MYOG*, *SOX8* and other TFs. Conversely, SOX8 is highly overexpressed in FP-RMS tumours, and SOX8 disruption leads to upregulation of *MYOD1* and *MYOG* in FP-RMS, suggesting a negative regulatory mechanism [9]. In conclusion, the FP-RMS CRC model includes feed forward (PAX3-FOXO1 and MYOD1, MYOG and other TFs) and negative feedback (SOX8) mechanisms [9,61].

In a more recent publication, Gryder and colleagues further dissect the CRC of FP-RMS and put forth a detailed mechanistic view of the chromosomal translocation that leads to hijacking of the *PAX3* promoter by *FOXO1* SE [62]. This group demonstrates that the SE of *FOXO1* interacts with smaller intergenic and intronic enhancers of *FOXO1* and *PAX3* promoter. In the stepwise developmental programme of skeletal muscle, PAX3 activates MYOD1 through *MYOD1* SE, but MYOD1 does not upregulate *PAX3*, and wild-type PAX3 enhancers are silent while MYOD1 and MYOG promote differentiation in late myogenesis [62]. By contrast, upon *FOXO1* SE translocation to regulate *PAX3* in FP-RMS, MYOD1, MYOG and MYCN can also bind to and drive this SE. This leads to the continuous expression of *PAX3-FOXO1* in late stages of myogenesis and halting of FP-RMS tumours in an undifferentiated state. These newly formed 'miswired' enhancer elements fuel the pathological diversion from normal skeletal muscle development in FP-RMS [62]. Table 1 summarizes examples of CRC TFs discussed in this section.

2.4. Renal cell carcinoma

Renal cell carcinoma (RCC) is a heterogeneous cancer accounting for 2% of all cancer cases [63]. Clear cell RCC (ccRCC) is the most common subtype of this disease (greater than 80% of all cases) and the main cause of RCC mortality. ccRCC harbours truncal mutations in the *VHL gene* (von Hippel-Lidau tumour suppressor) implicated in activation of TFs such as HIF1α and HIF2α that are involved in angiogenesis, metabolism and cell death [64]. However, consistent with the Knudsen's two-hit genetic alteration hypothesis, the addition of a second genetic alteration in mTOR pathways or chromatin modifiers is also required for induction of ccRCC [65]. In a recent study, PAX8, a cell-autonomous transcriptional activator, was identified as a potential CRC oncogenic driver in RCC, which may be independent of *VHL* alteration status [37]. *PAX8* knockdown in an array of RCC cell lines revealed a network of over 460 genes including those involved in metabolism, kidney cell fate, proliferation and the process of tumourigenesis (e.g. kidney-specific cadherins, claudins and cell cycle genes) under PAX8 regulation. One key difference between PAX8 regulation of metabolic genes compared with its other targets was the prevalence of H3K27ac. Specifically, cell cycle and metabolic pathway genes gained H3K27ac marks indicating that they were enhancer-regulated by PAX8, rather than promoter-regulated [37]. An example of a PAX8 target gene (and also HIF) is ferroxidase ceruloplasmin (CP), implicated in the iron-metabolic pathway in RCC tumourigenesis [37]. CP is also a marker of refractory disease and low survival in RCC patients in addition to being a predictor of PAX8 activity [37]. Table 1 summarizes examples of CRC TFs discussed in this section.

2.5. Liposarcoma

Liposarcomas (LPSs), or soft tissue sarcomas, are mesenchymal tumours that account for 20% of adult sarcomas [66]. Somatic abnormalities in LPS tumours comprise overexpression of CDK4 and MDM2, and 12q13–15 amplification [67]. Four LPS subtypes have been identified; well-differentiated (WDLPS), myxoid (MLPS), pleomorphic (PLPS) and de-differentiated (DDLPS), the latter three comprising most high-grade cases; PLPS and DDLPS mainly lead to disease relapse post-treatment, while MLPS displays better prognosis [68].

Charting H3K27ac modifications of LPS (DDLPS and MLPS) cell lines and primary tissue, mesenchymal stem cells and mature adipocytes, revealed that some SEs are retained from the adipogenesis programme (e.g. *FOSL2*). By contrast, SEs of definitive adipocyte genes are ablated (e.g. *CEBPA* and *PPARG*) while there is *de novo* establishment of SEs related to genes associated with transformation (e.g. *MYC*, *CDK6* and *JUN*) [38]. In these LPS samples, the SEs preferentially used are those associated with tumourigenesis, including cell migration, angiogenesis and other developmental processes [38]. Finally, a low-to-moderate overlap was observed between DDLPS and MLPS SEs in primary tissue and cell lines [38].

The defining factor in the MLPS CRC is a fusion oncogene resulting from the t(12;16)(q13;p11) translocation, forming a hallmark MLPS FUS-DDIT3 fusion which functions as a TF [69,70]. FUS-DDIT3 is disproportionately distributed in the genome, especially in SE regions contributing to deregulated gene expression and an aberrant epigenetic landscape. One interesting observation in this subtype was transcriptional addiction owing to preferential SE association with genes regulating RNA-Pol2 activity. Consistent with this, close to 9% of FUS-DDIT3 bound to promoters with high RNA-Pol2 activity [38]. When present, a double H3K27ac and FUS-DDIT3 mark led to high basal expression levels (e.g. *FST* and *IL8*), displaying its potential for corruption of epigenetic landscapes. A known group of interactors with histone acetylation marks of SE regions are bromodomain and extra terminal domain proteins (BET) [71]. Consistent with the notion that oncogenic fusion TFs hijack BET proteins to activate malignant transformation, substantial co-localization and co-operation between FUS-DDIT3 and the BET protein BRD4 has been detected in MLPS [11,38]

CRCs associated with DDLPS comprise FOSL2, MYC and RUNX1, whose maintenance is dependent on BET proteins. Marked co-occupancy of RUNX1 and FOSL2 activates a network of targets involved in the pathogenesis of liposarcoma and malignant growth [38]. Specifically, FOSL2 and RUNX1 proteins co-occupy the SE regions of all described CRC TFs in this LPS subtype. These genes collectively maintain the expression of SNAI2, indicated in EMT and proliferative capacity, and a potential prognostic marker for this subtype. Higher SNAI2 is also linked to shorter disease-free survival (DFS) in DDLPS patients [38]. Finally, demonstrating the dependency of the DDPLS CRC on BRD4, depletion of BRD4 attenuated distant metastasis [38]. Table 1 summarizes examples of CRC TFs discussed in this section.

2.6. Prostate cancer

Prostate cancer is one of the major causes of cancer-related deaths in men [72]. The androgen receptor (AR) dictates the transcriptional output that promotes proliferation and survival of prostate cancer cells. Studies focused on dissecting the mechanisms of AR-centred prostate cancer development reveal that AR not only regulates gene expression but also regulates higher-order chromatin configuration [73]. More specifically, a study [39] identified that 55% of AR binding sites function as anchors that mediate duplex and complex AR-associated chromatin interactions (AR_{anchor}), while the remaining 45% did not participate in chromatin interaction (AR_{alone}). There was a two-fold enrichment of androgen upregulated genes in AR_{anchor} regions compared with AR_{alone} regions, which highlights that long-range chromatin looping may be pivotal to AR regulatory functions [39].

TFs can interact with nuclear hormone receptors such as the AR to govern different aspects of transcription and chromatin regulation [74]. A recurrent fusion gene in prostate cancers, ERG (erythroblast transformation-specific related gene), was shown to interact and collaborate with AR through chromatin looping [73,74]. The ERG interactome, including ERG-associated long-range chromatin, is a collaborative component of higher-order AR-associated chromatin structure and is involved in co-regulating subtypes of AR target genes in prostate cancer. For instance, this study detected intertwined ERG-associated and AR-associated chromatin loops in relation to genes or gene clusters such as *FKBP5*, *VCL*, *KLK* family, *EAF2* and *SLC15A2-ILDR1* [39].

AR and ERG co-bind to regulatory sites associated with long-range chromatin interactions ($AR^{+}ERG^{+}_{anchor}$). These sites have been shown to be associated with enhancer activity,

TF binding motifs and bi-directional transcription [39]. Further, these AR and ERG-associated highly connected hubs co-localized with sites for binding of epigenetic regulators/histone remodelling factors and lncRNAs [39]. With regard to co-localization of epigenetic regulators/histone remodelling factors with distinct AR-ERG transcriptional network, three distinct genomic signatures were identified: (i) FOXA1, EZH2 and HDAC3 that are enriched with $AR^{+}ERG^{+}_{anchor}$ sites; (ii) HDAC1, BRD2, BRD3 and BRD4 that are enriched with $AR^{-}ERG^{+}_{anchor}$ and $AR^{-}ERG^{+}_{alone}$ (ERG in the absence of AR); and (iii) POLR2A, HDAC2 and GAPBPA that are enriched with AR looping but not $AR^{+}ERG^{+}_{alone}$ and $AR^{+}ERG^{-}_{alone}$ [39].

With respect to lncRNAs, one potential function of AR and ERG chromatin looping may be to allow interactions between lncRNA and its target gene. For instance, manipulating three lncRNAs identified in association with the *PMEPA1* locus (*PCAT43*, *PCAT61* and *PCAT76*) led to a reduction in androgen-triggered expression of the gene [39]. One other example of the clinical relevance of AR and ERG chromatin loops is the link detected between a prostate cancer GWAS SNP, rs9364554, located in the intron of *SLC22A3* within an AR and ERG loop anchor. This loop also connects this SNP with *SLC22A2* in the vicinity [39]. Table 1 summarizes examples of CRC TFs discussed in this section.

2.7. Gastrointestinal stromal tumour

Gastrointestinal stromal tumour (GIST) is a common soft tissue sarcoma, originating from interstitial cells of Cajal (ICC) [75]. The ICC lineage is reliant on KIT and ETV1 for specification and survival, whereby KIT and ETV1 function as signalling and lineage-specific regulators, respectively [75,76]. During development, the transcriptional input required for ICC lineage specification constitutes KIT activation by KIT ligand and consequent MAPK-mediated stabilization of ETV1 protein, establishing lineage specification [75]. In the pathological context, mutant KIT stabilizes ETV1 (through aberrant MAPK signalling activation), while in turn, ETV1 promotes mutant KIT expression, forming a divergent positive feedback loop fuelling the process of tumourigenesis [40].

FOXF1, a member of the fork-head family of transcription factors, is specifically expressed in GIST and directly regulates the transcription of *KIT* and *ETV1*. In turn, FOXF1 and ETV1 both regulate KIT, although FOXF1 regulation of KIT is significantly stronger owing to the regulation of both chromatin accessibility and the ETV1 cistrome [40]. This evidence may support the pre-existence of this regulatory pattern between KIT and FOXF1 in non-oncogenic ICC development, highlighting similarities between physiological and pathological development.

FOXF1 also co-localizes with ETV1 to regulate ICC/GIST lineage-specific gene expression by maintaining open chromatin structure and enhancers, as well as the recruitment of ETV1 to lineage-specific enhancers. Examples of ETV1-dependent ICC/GIST lineage-specific gene networks regulated by FOXF1 include *DUSP6*, *GPR20* and *ANO1* [40].

With respect to FOXF1 regulation, KIT or MAPK pathway perturbations do not significantly affect the expression of *FOXF1*, placing it at the top of a regulatory hierarchy for GIST. Finally, FOXF1 is required for GIST cell cycle progression, tumour growth and maintenance [40]. Table 1 summarizes examples of CRC TFs discussed in this section.

2.8. Medulloblastoma

Medulloblastoma, a malignant paediatric brain tumour arising from the cerebellum, medulla and brain stem, is categorized into four clinically and biologically distinct subgroups [77]. These four core subgroups, WNT, SHH, group 3 and group 4, are classified based on their inherent differential and discriminatory transcriptional profiles. The WNT and SHH subgroups are named based on the activity of the respective pathways, and groups 3 and 4 display regulatory similarities [78] but present diverse phenotypes and express GABAergic and glutaminergic cell-type characteristics, respectively [21,77]. In addition to somatic alterations in driver genes such as *MYC* (group 3), *KDM6A* (group 4) and *GFI1*/*GFI1B* (group 3 and 4) [21,77,79], epigenetic modulation may influence transcriptional programming specific to subgroups [80].

The computational reconstruction of SE and enhancer mapping for 28 medulloblastoma primary tissue has been used to dissect differential group 3 and 4 CRCs [21]. This mapping approach identified large SEs associated with cerebellum-specific TFs, *ZIC1* and *ZIC4*, and SEs associated with medulloblastoma driver genes and epigenetic modulators, such as *GLI2*, *MYC* and *OTX2* [21]. On a subgroup level, SEs were then inferred to regulate *ALK* in the WNT group, *SMO* and *NTRK3* in the SHH group, *LMO1*, *LMO2* and *MYC* in group 3, and *ETV4* and *PAX5* in group 4 [21]. This group-specific SE allocation was based on an unbiased hierarchical clustering strategy of SEs across the samples analysed. One key observation in the study was that SE patterns observed differed substantially between medulloblastoma primary tissue or cell lines highlighting regulatory and CRC component dissimilarities [21]. This study also identified core TFs implicated in establishing medulloblastoma group identity including HLX (group 3), LMX1A (group 4) and LHX2 (shared between groups 3 and 4), providing some evidence towards the cell-of-origin of these disease groups [21]. In terms of functional pathway enrichment, TGFβ signalling and neuronal transcriptional regulators were enriched in groups 3 and 4, respectively [21]. Table 1 summarizes examples of CRC TFs discussed in this section.

2.9. Chronic lymphocytic leukaemia

Chronic lymphocytic leukaemia (CLL) is a highly heterogeneous B-cell haematological malignancy with low cure rates. A spectrum of genomic alterations in this malignancy have been identified, including segmental chromosomal alterations, copy number alterations and somatic nucleotide alterations, while 13q deletion is the most recurrent alteration [81,82]. The CLL-specific CRC is centred on PAX5, a TF that promotes lymphomagenesis by activating signalling pathways indicated in B-cell signalling, and the knockdown of this gene results in dramatic effects on B-cell proliferation and development [41,83].

In a study aimed at dissecting CRCs in primary CLL and normal B cells (NBCs), SEs with exceptionally high H3K27ac marks (42% of all H3K27ac marks globally) were discovered in proximity to genes involved in CLL pathobiology, including *CXCR4*, *CD74*, *PAX5*, *CD5*, *KRAS* and *BCL2* [41]. This

high proportion of H3K27ac at these few loci of total global H3K27ac activity concomitant with open chromatin structure (tested by ATAC-seq) demonstrates the dominance of these SEs in regulating transcriptional output. For instance, the SE of the *BCL2* gene that is usually upregulated in CLL, open chromatin structure and broad H3K27ac signals were detected [41]. The SE of *CTLA4*, encoding a T-cell inhibitory checkpoint effector, also displayed strong H3K27ac signals. The NBC samples used in this study showed 230 SEs, including SEs proximal to *BACH2* and *BANK1*, known to play roles in lymphoma suppression [41,84]. Further, despite samples displaying substantial heterogeneity, a core of large SEs displayed regulatory conservation among a subset of the CLL patient samples in loci pertinent to *KRAS*, *CD5*, *PAX5*, *CXCR4*, *BCL2* and *CD74* [41]. Finally, this study defines an enhancer-based CRC analysis system. Specifically, for TFs associated with top-ranked enhancers, inward TF enhancer binding by other TFs and outward binding of the TF of interest to their extended enhancer network were assessed. This information was processed to describe 'cliques' of auto-regulatory TFs [41]. At least four representative cliques were defined: CLL-2, CLL-3, CLL-8 and CLL-11. For instance, TFs constituting the CLL-2 clique include PAX5, ETV6, TCF3, IRF2, MEF2D, ELF1, KLF13, JUND, FOXP1, IRF1 and IRF8 [41]. Highly connected CLL and NBC TFs across samples comprised PAX5 and the IRF family in addition to FOXP1, RARA and ETS1 [41]. Table 1 summarizes examples of CRC TFs discussed in this section.

2.10. T-cell acute lymphoblastic leukaemia

For T-cell acute lymphoblastic leukaemia (T-ALL), malignant transformation gives rise to leukaemic cells owing to deregulated thymic differentiation programmes [85]. The oncogenic TF, TAL1, is crucially involved in the pathogenesis of T-ALL cases and has been shown to collaborate with other TFs to form a CRC. This CRC comprises TAL1, HEB, E2A, LMO1/2, GATA3 and RUNX1 in T-ALL representative cell lines, such as Jurkat and CCRF-CEM [42]. A high coincidence of genomic site occupation was observed in this study between TAL1 and other CRC TFs including LMO1/2, GATA3 and RUNX1. In these two cell lines, three different classes of regulatory elements were identified: group 1 (concordant enrichment for TAL1 complexes), group 2 (mainly GATA3 occupation) and group 3 (mainly RUNX1 occupation). In terms of the presence of identified binding motifs, these were, for group 1, E-box, GATA, RUNX and ETS, for group 2 GATA and ETS motifs, and for group 3 RUNX, ETS and SP1 [42].

In summary, TAL1 forms an auto-regulatory loop with GATA3 and RUNX1, and they occupy regulatory regions of their own and each other's genes. TAL1 initiates this auto-regulatory loop, and the sustained upregulation of GATA3 and RUNX1 by TAL1 may contribute to reinforcement of the malignant programme in T-ALL [42]. Further, TAL1 positively regulates the expression of a network of target genes in collaboration with GATA3 and RUNX1 [42].

Target genes of TAL1 include *TRIB2* and *MYB* whereby the former regulates cell survival in TAL1-positive T-ALL cells, while the latter is a transcriptional regulator driving normal and malignant blood haematopoiesis [86]. *MYB* is induced by TAL1 and in turn, MYB co-regulates a subset of TAL1 target genes, stabilizes and reinforces the TAL1 oncogenic programme [42]. One example for collaboration between TAL1 and MYB in TAL1-positive T-ALL cells is that the enhancer region of TAL1 can be targeted by numerous somatic alterations which then form new MYB binding sites and SEs, effectively extending the outreach of MYB [18]. An example of negative and positive regulation in T-ALL is the TAL1, HEB and H2A regulatory network. TAL1, HEB and H2A coordinately regulate target genes. Of these target genes, a subset is directly activated by TAL1 but repressed by HEB and H2A [42]. Table 1 summarizes examples of CRC TFs discussed in this section.

3. CRCs and drug development

The dissection of regulatory networks associated with cell identity in cancer facilitates a better understanding of the malignancy and the identification of appropriate treatment strategies. CRCs provide a framework for the identification and potential targeting of oncogenic CRC TFs, transcriptional co-activators, SEs and SE-associated co-activators and modulators as justifiable avenues of targeting. One example of targeting master regulator TFs for therapeutic gain is in GIST. This cancer is highly resistant to standard chemotherapy, and is instead sensitive to specific targeting of KIT and ETV1 lineage-specific CRC TFs [87,88]. Further, CRC TFs recruit acetylation writers such as CBP/p300, readers such as BRD4 and erasers such as HDACs and other factors to construct SEs [8,22]. BRD4 and related proteins have been shown to occupy large numbers of enhancers, especially SEs [11,15]. Due to this association, SEs may be sensitive to drugs that target BET domain regulators and kinases involved in transcription [15,89]. Despite the broad presence of BET proteins across thousands of enhancers, inhibition of these proteins (for instance the inhibition of BRD4 by the BET-bromodomain inhibitor JQ1), has led to specific targeting in multiple cancers, revealing cancer dependencies. In multiple myeloma, JQ1 treatment led to specific MYC inhibition [15] (figure 2), while in CLL, BET inhibition led to the downregulation of multiple survival pathways involved in CLL biology [90]. This pattern was also observed in diffuse large B-cell lymphoma (DLBCL), in which SEs of oncogenic and lineage-specific CRCs showed particular sensitivity to BET inhibition [11].

In addition to gene or gene network targeting, BET protein inhibition may be explored to sensitize cases of relapse and treatment resistance. For instance, in solid tumours such as LPS, targeting BET proteins using ARV-825, a BET protein degrader, can provide advantages in overcoming trabectedin resistance [38]. In terms of cellular effects, BET protein inhibition and depletion mainly triggers apoptosis or cytotoxic effects in cancers, including osteosarcomas and breast cancer [91,92].

One other outcome of chemical targeting of SEs is to understand SE driven transcriptional addiction in cancers. In multiple myeloma, JQ1 treatment more dramatically affects SEs and SE-associated genes compared with typical enhancer-associated genes [15]. Cancer addiction to CR transcription has been described in RMS, in which the PAX3-FOXO1 fusion protein activates SEs to activate the expression of other CR TFs in a feed-forward manner, leading to high levels of CR TF expression [22]. Consistent with transcriptional addiction, the selective disruption of CR transcription

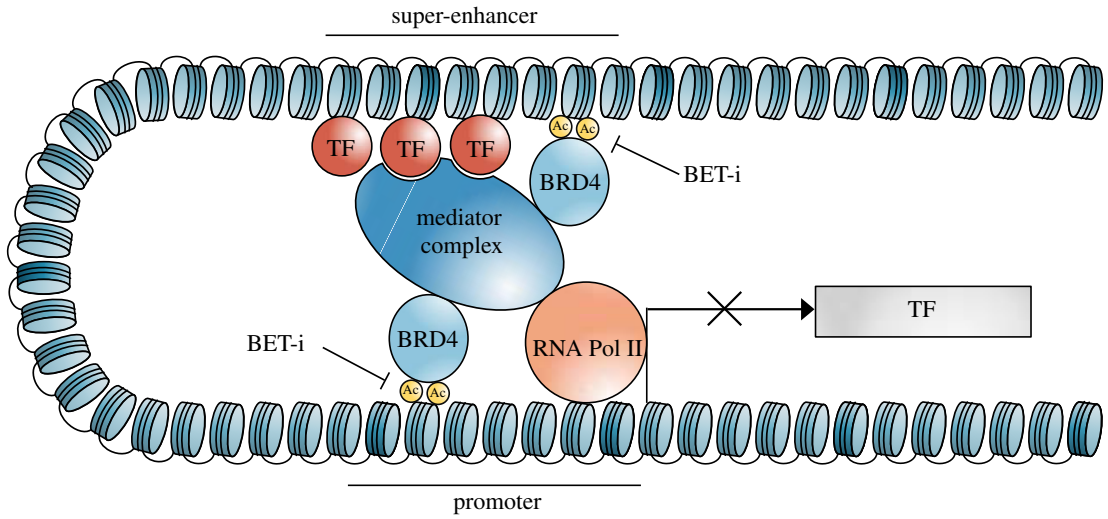


Figure 2. BET inhibitor treatment represses transcription of super enhancer-associated transcription factors. BET proteins (including BRD4) regulate chromatin and RNA polymerase accessibility to the gene of interest. BET inhibitors such as JQ1 can induce disruption of SEs and specific transcription elongation defects and inhibition by displacing BRD4.

was achieved by targeting the acetylation axis in this cancer [22]. Specifically, this study showed that co-inhibition of HDAC1, HDAC2 and HDAC3 halts CR transcription by interfering with chromatin accessibility and looping [22]. In conclusion, understanding the dependency and mechanistic connections between BET proteins and deregulated programmes and enhancer states can provide avenues for target identification and therapeutic gain.

4. Similarities between CRC models in the wider context of human cancer

The CRC TFs identified in this study, although displaying specific functions in each cancer's CRC, may indeed be involved in gene regulatory networks in a spectrum of other human cancer cell lines and primary tissues. The dbCoRC database permits the collation of information concerning cell or tissue expression of a given CRC TF, upstream and downstream targets of this TF within the CRC model, SE genomic coordinates and the number of TF binding sites within the SE of the targets (CRC TFs) [19]. Here, we have used this tool to further study the CRC TFs indicated in the 10 cancer types discussed in this review in other cell lines. Table 2 outlines reviewed CRC TFs indicated in other human cell lines and primary tissue, and the CRC model formed. For instance, FOXP1 and ERG reviewed in the context of CLL and prostate cancers, respectively, are both indicated in the CRC model of a colorectal cancer cell line, COLO320 (ASCL2, DBP, ERG, FOXG1, FOXP1, MEIS1, OSR1, SOX5, SP1, TBX2, TEAD1, TFAP2C and TFAP4). Another example is GATA3, which has been reviewed in this article in NB and T-ALL, and has also been identified in the regulatory networks of a breast cancer cell line (ZR-75-1) [19,31,42]. This regulatory network comprises TFs such as: EHF, FOXA1, GATA3, HES1, MEF2D, NFIB, NR2F2, OSR2, PATZ1, RARA, SP2, SP3, SPDEF, SREBF1, YY1 and TGIF1 (figure 3*a*). The CRC model proposed by dbCoRC for GATA3 in breast cancer was further processed using DisGeNet to test the association of these TFs with other cancers and other diseases (figure 3*b*) [93,94]. Using this programme and without correction for multiple testing, strong associations with several cancers were identified. Each link represents the number of *overlapping* genes annotated to each term, and size represents the number of genes annotated to each term. These data highlight the importance of understanding and comparing TFs across a wider spectrum of cancer cell lines and primary tissue, with the objective of the discovery of overlapping and non-overlapping functions and mechanisms.

5. Methodologies that facilitate the understanding of CRCs

Next-generation technologies have allowed shifting from inter-patient tumour variability to the precise characterization of intra-tumour genetic, genomic and transcriptional heterogeneity via multi-regional bulk tissue NGS. Emerging single-cell transcriptomics, coupled with NGS, allow novel strategies for therapeutic response prediction and drug development. The regulatory mechanisms that govern the transcriptome and the expression of these regulatory circuits are now being investigated using WGS to identify non-coding mutations and chromatin profile using ChIP-seq (chromatin immunoprecipitation followed by sequencing), 4-C (circulized chromatin conformation capture), ChIA-PET (chromatin interaction analysis with paired-end tags) and ATAC-seq (assay for transposase-accessible chromatin followed by sequencing) [95,96]. Understanding regulatory networks at single-cell resolution has empowered efforts to decipher cancer heterogeneity, differential resistance to therapy patterns and hierarchical classification, for instance, in breast cancer [97]. Here, we briefly elaborate on each method.

ChIP-seq is a technique allows the detection of TF binding profiles and histone modifications, including the H3K27ac marks that signify SEs. The challenge with this technique is obtaining a highly specific antibody [8].

4C-seq is an update of the chromosome conformation capture (3C) coupled to sequencing (Hi-C) method that quantifies contact frequencies of DNA based on nuclear proximity,

Table 2. Summary of the CRC network data extracted from dbCoRC database for CRC TFs discussed in this review. CRC TFs discussed in this study were investigated using dbCoRC database to identify the differential utility of these TFs in CRCs models of other human cancer cell lines and primary tissue. The example provided for the implication of the TF in other cancer cell line or primary tissue represents one of many examples provided by this database.

CRC TF/ malignancy	other cancer cell lines or primary tissue	examples of upstream/downstream TFs within the CRC model in this cell line or primary tissue
PHOX2B/ NB	NCI-H82 (SCLC)	OTX2, SREBF1, TEAD1, MYC, NHLH1, NR2F6, PHOX2B
GATA3/NB and T-ALL	ZR-75-1 (breast carcinoma)	EHF, FOXA1, GATA3, HES1, MEF2D, NFIB, NR2F2, OSR2, PATZ1, RARA, SP2, SP3, SPDEF, SREBF1, YY1, TGIF1
SOX2/GBM	NCI-H69 (SCLC)	BARHL1, DLX1, ETS1, FOXA1, SOX2, FOXG1, INSM1, KLF13, KLF7, MSX2, NR2F1, SP8, TCF4, TEAD1
MYOD1/RMS	RH18 (RMS)	ARID3A, FOXL1, GLI1, GLI3, HOXC9, IRF1, MAFK, MYOD1, RARA, RXRA, SMAD3, TBX1, TEAD3, VDR
MYOG/RMS	RD (RMS)	ETV4, GLI3, HOXC10, HOXC9, HOXD8, KLF7, MYOD1, MYOG, RUNX1, SMAD3, SOX8, TCF7L2, ZNF219
MYC/LPS	NCI-H82 (SCLC)	MYC, NHLH1, NR2F6, OTX2, PHOX2B, SREBF1, TEAD1
RUNX1/LPS and T-ALL	T20020720 (gastric cancer)	EHF, ELF3, ETS2, IRF1, IRF2, KLF13, KLF5, MAFF, MEIS1, NR4A1, PRDM1, RREB1, RXRA, SMAD3, SOX13, TCF7L2, RUNX1
ERG/prostate	COLO320 (colorectal cancer)	ASCL2, DBP, ERG, FOXG1, FOXP1, MEIS1, OSR1, SOX5, SP1, TBX2, TEAD1, TFAP4, TFAP2C
PAX5/CLL	SU-DHL-6 (diffuse large B-cell lymphoma)	ARID5B, CUX2, ELF1, MAX, PAX5, SMAD3
ETV6/CLL	T2000085 (gastric cancer)	BCL6, BHLHE40, ETS1, ETV6, GLIS3, HIVEP2, IKZF1, IRF2, KLF7, MEF2D, MEIS1, NR2F2, RARA, RREB1, RUNX1, SMAD3, ZBTB16
IRF2/CLL	HCC1954 (breast cancer)	ELF3, FOXI1, HES1, IKZF2, IRF2, NFIA, PBX1, PITX1, SP3, STAT4, TFAP2A, TP63
ELF1/CLL	COLO205 (colorectal cancer)	ASCL2, BARX2, BHLHE40, DLX2, EHF, ELF1, ELF3, FOS, FOXB1, HES1, HNF1B, IRF1, IRF8, KLF5, MYB, PDX1, PITX1, RREB1, RUNX1, RUNX3, SMAD3, SREBF1, TCF7, TCF7L2, TEAD1
KLF13/CLL	T2001206 (gastric cancer)	BHLHE40, ELF3, ETS1, ETS2, ETV6, HIF1A, IRF1, IRF2, KLF5, KLF13, MEIS1, PRDM1, RREB1, RUNX1, SMAD3, TCF7L2, TGIF1
FOXP1/CLL	COLO320 (colorectal cancer)	ASCL2, DBP, ERG, FOXG1, FOXP1, MEIS1, OSR1, SOX5, SP1, TBX2, TEAD1, TFAP2C, TFAP4
NFATC1/CLL	HBL1 (diffuse large B-cell lymphoma)	BACH2, EBF1, ETS1, ETV6, FOXP1, HES1, IKZF1, IRF2, IRF4, IRF8, MAX, MEF2A, MEF2D, NFATC1, NR3C1, PAX5, POU2F2, RORA, RUNX1, TCF4, TBX15, TFEB
KLF12/CLL	COLO741 (colorectal cancer)	EGR3, EN2, ETS1, KLF12, NR4A1, NR4A2, PKNOX2, RARA, RREB1, SMAD3, SNAI2, SP1, SREBF1, TBX2, TEAD1, TGIF1
JUN/CLL	MiaPaca2 (pancreatic adenocarcinoma)	EHF, HOXB6, JUN, MYBL1, MYC, NR2F2, NR5A2, RXRA, SHOX2, SMAD3, SREBF1, TBX4, TP63, WT1

and reveals chromatin folding and configuration patterns [98]. 4C-seq takes into account domains of contact and inter-domain contact of a specific genomic site within genome sequences [99]. The main limitation of 4C is technical biases due to coverage of *cis* and *trans* chromosome interactions and the use of restriction enzymes [100,101]. ChIA-PET detects chromatin interactions associated with a protein of interest. This method is unbiased and relies on the premise that proximal DNA sequences from the same cross-linked molecular complex may be ligated, offering enhanced resolution and throughput compared with previous techniques [100]. The limitations of ChIA-PET include the requirement for substantial starting material due to the sequence of experimental steps. An improved adaptation of this method is proximity ligation-assisted ChIP-seq (PLAC-seq), which features shifting forward of the ligation step. Briefly, in this method, *in situ* proximity ligation is performed prior to lysis of the nuclei, significantly reducing the required input material and improving the efficacy and accuracy over ChIA-PET [102]. Another improved method of detecting chromatin conformation mediated by a protein of interest that addresses limitations of ChIA-PET is HiChIP. This method also relies on *in situ* establishment of DNA contacts prior to lysis of nuclei. Subsequently, ChIP and on-bead library generation is carried out followed by paired-end sequencing, revealing the long-range interactome of the protein of interest [103]. A significant drawback of HiChIP is the effect of sequencing depth on the accuracy of detected interactions. Gryder and colleagues address this drawback by introducing AqUa-HiChIP [104]. This method circumvents the limitation of HiChIP by absolute quantification of chromatin interactions. Briefly, this method relies on a previously defined ratio of formalin-fixed nuclei of two different origins (for instance mouse versus human nuclei). The nuclei are lysed, and upon incorporation and ligation of biotin-dATP, shearing is performed. Subsequently, ChIP,

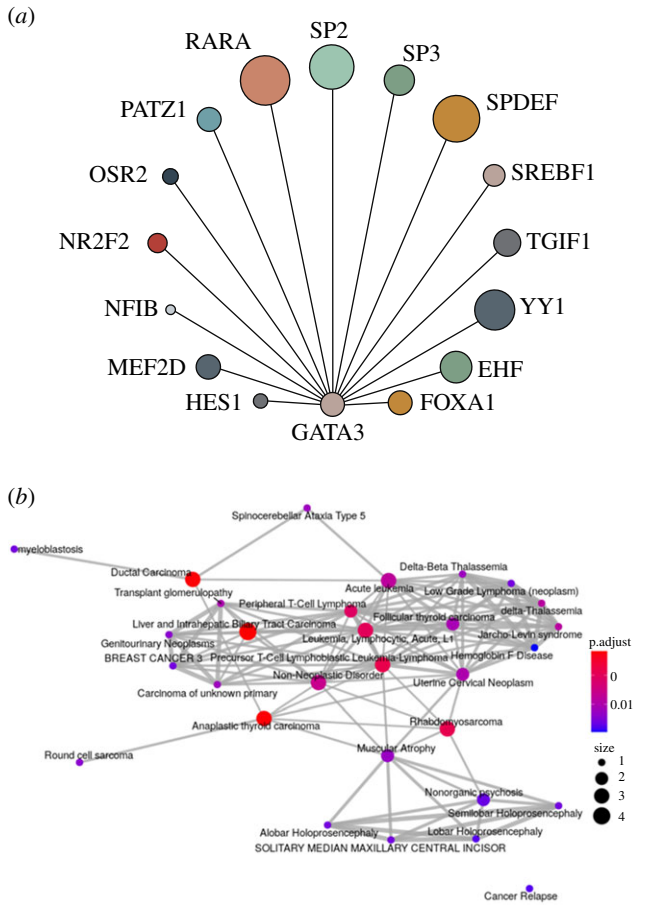


Figure 3. (*a*) GATA3 is indicated in the CRC model of the ZR-75-1 breast cancer cell line. CRC model output from dbCoRC for GATA3 in the ZR-75-1 breast cancer cell line involving GATA3 regulation of 15 target TFs (by binding to their SEs). The circumference of each target TF circle is proportionate to the number of GATA3 binding sites identified in SEs of the gene encoding this target TF. (*b*) Using DisGeNet without multiple testing corrections, the CRC model TFs in (*a*) show an association with several cancers and other diseases. Each link represents the number of overlapping genes annotated to each term, and size represents the number of genes annotated to each term.

biotin capture and paired-end sequencing are performed [104]. Human chromatin interactions are then normalized to those of the mouse genome on the grounds of paired-end tag counts, allowing more accurate quantification of these interactions. Alongside the experimental method, this group also provides a streamlined bioinformatics analysis platform coupled to this method [104].

ATAC-seq assays the transposase accessibility of chromatin coupled with next-generation sequencing. It relies on the insertion of sequencing linkers by a hyperactive Tn5 transposase enzyme. Sequencing of the linker attached to reads reveals regions of chromatin accessibility and offers higher sensitivity compared with other techniques such as DNAse-seq. Limitations in streamlined bioinformatics analysis pipelines may be a challenge with this technique [105]. Finally, single-cell-resolution ATAC-seq can inform areas of chromatin accessibility and shed light on developmental processes [106].

6. Conclusion

This review summarizes CRC TF members associated with SEs in a range of liquid and solid cancers. CRC TFs create and maintain cell-type specific regulatory programmes and define cell identity, a process that is deregulated in many cancer subtypes. Specific TFs play important roles in forming CRC networks in several types of cancer cell lines and primary tissues, suggesting similar yet divergent mechanisms and players involved in regulatory processes. Reconstruction of CRCs in cancer cell lines and tissue, obtained by leveraging genomic technologies, will facilitate the understanding of deregulation of biological processes in carcinogenesis and support the reconstruction of a blueprint pertaining to the identity of a cancer. Consistent with this, transcriptional addiction is emerging as an important novel drug vulnerability in cancers. Therefore, understanding components of CRCs, associated proteins and regulators can provide opportunities for targeting of these components for therapeutic advantage.

Data accessibility. This article has no additional data.

Authors' contribution. L.J., L.T., R.M.T. and J.P. collected data and wrote the paper. J.A.W., G.V.G. and S.D.T. wrote the paper. R.M.T. designed the figures.

Competing interests. We declare we have no competing interests.

Funding. There was no funding associated with this article.

Acknowledgements. G.V.G. acknowledges support from the NIHR Birmingham ECMC, NIHR Birmingham SRMRC, Nanocommons Horizon 2020-EU (731032), the NIHR Birmingham Biomedical Research Centre and the MRC HDR UK (HDRUK/CFC/01), an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the Medical Research Council or the Department of Health.

References

1. Pereira JD, Sansom SN, Smith J, Dobenecker MW, Tarakhovsky A, Livesey FJ. 2010 Ezh2, the histone methyltransferase of PRC2, regulates the balance between self-renewal and differentiation in the cerebral cortex. *Proc. Natl Acad. Sci. USA* **107**, 15 957–15 962. (doi:10.1073/pnas.1002530107)
2. Sansom SN, Griffiths DS, Faedo A, Kleinjan DJ, Ruan Y, Smith J, van Heyningen V, Rubenstein JL, Livesey FJ. 2009 The level of the transcription factor Pax6 is essential for controlling the balance between neural stem cell self-renewal and neurogenesis. *PLoS Genet.* **5**, e1000511. (doi:10.1371/journal.pgen.1000511)
3. Tuoc TC, Boretius S, Sansom SN, Pitulescu M, Frahm J, Livesey FJ, Stoykova A. 2013 Chromatin regulation by BAF170 controls cerebral cortical size and thickness. *Dev. Cell.* **25**, 256–269. (doi:10.1016/j.devcel.2013.04.005)
4. Raja DA *et al.* 2020 Histone variant dictates fate biasing of neural crest cells to melanocyte lineage. *Development* **147**, dev182576. (doi:10.1242/dev.182576)
5. Kuznetsov JN, Aguero TH, Owens DA, Kurtenbach S, Field MG, Durante MA, Rodriguez DA, King ML, Harbour JW. 2019 BAP1 regulates epigenetic switch from pluripotency to differentiation in developmental lineages giving rise to BAP1-mutant

cancers. *Sci. Adv.* **5**, eaax1738. (doi:10.1126/sciadv.aax1738)

6. Saint-André V, Federation AJ, Lin CY, Abraham BJ, Reddy J, Lee TI, Bradner JE, Young R. 2016 Models of human core transcriptional regulatory circuitries. *Genome Res.* **26**, 385–396. (doi:10.1101/gr.197590.115)
7. Hnisz D, Schuijers J, Lin CY, Weintraub AS, Abraham BJ, Lee TI, Bradner JE, Young RA. 2015 Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. *Mol. Cell.* **58**, 362–370. (doi:10.1016/j.molcel.2015.02.014)
8. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, Hoke HA, Young RA. 2013 Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947. (doi:10.1016/j.cell.2013.09.053)
9. Gryder BE *et al.* 2019 Histone hyperacetylation disrupts core gene regulatory architecture in rhabdomyosarcoma. *Nat. Genet.* **51**, 1714–1722. (doi:10.1038/s41588-019-0534-4)
10. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA. 2013 Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319. (doi:10.1016/j.cell.2013.03.035)
11. Chapuy B *et al.* 2013 Discovery and characterization of super-enhancer-associated dependencies in diffuse large B cell lymphoma. *Cancer Cell.* **24**, 777–790. (doi:10.1016/j.ccr.2013.11.003)
12. Boeva V *et al.* 2017 Heterogeneity of neuroblastoma cell identity defined by transcriptional circuitries. *Nat. Genet.* **49**, 1408–1413. (doi:10.1038/ng.3921)
13. Young R. 2011 Control of the embryonic stem cell state. *Cell* **144**, 940–954. (doi:10.1016/j.cell.2011.01.032)
14. Boyer LA *et al.* 2005 Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947–956. (doi:10.1016/j.cell.2005.08.020)
15. Lovén J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, Bradner JE, Lee TI, Young RA. 2013 Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320–334. (doi:10.1016/j.cell.2013.03.036)
16. Takahashi K, Yamanaka S. 2006 Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676. (doi:10.1016/j.cell.2006.07.024)
17. Oldridge DA *et al.* 2015 Genetic predisposition to neuroblastoma mediated by a LMO1 super-enhancer polymorphism. *Nature* **528**, 418. (doi:10.1038/nature15540)
18. Mansour MR *et al.* 2014 An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346**, 1373–1377. (doi:10.1126/science.1259037)
19. Huang M, Chen Y, Yang M, Guo A, Xu Y, Xu L, Koeffler HP. 2018 dbCoRC: a database of core transcriptional regulatory circuitries modeled by H3K27ac ChIP-seq signals. *Nucleic Acids Res.* **46**, D71–D77. (doi:10.1093/nar/gkx796)
20. Khan A, Zhang X. 2016 dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Res.* **44**, D164–D171. (doi:10.1093/nar/gkv1002)
21. Lin CY *et al.* 2016 Active medulloblastoma enhancers reveal subgroup-specific cellular origins. *Nature* **530**, 57–62. (doi:10.1038/nature16546)
22. Gryder BE *et al.* 2019 Chemical genomics reveals histone deacetylases are required for core regulatory transcription. *Nat. Commun.* **10**, 3004. (doi:10.1038/s41467-019-11046-7)
23. Marris JM. 2010 Recent advances in neuroblastoma. *N. Engl. J. Med.* **362**, 2202–2211. (doi:10.1056/NEJMra0804577)
24. van Groningen T *et al.* 2019 A NOTCH feed-forward loop drives reprogramming from adrenergic to mesenchymal state in neuroblastoma. *Nat. Commun.* **10**, 1530. (doi:10.1038/s41467-019-09470-w)
25. van Groningen T *et al.* 2017 Neuroblastoma is composed of two super-enhancer-associated differentiation states. *Nat. Genet.* **49**, 1261–1266. (doi:10.1038/ng.3899)
26. Fryer CJ, Lamar E, Turbachova I, Kintner C, Jones KA. 2002 Mastermind mediates chromatin-specific transcription and turnover of the Notch enhancer complex. *Genes Dev.* **16**, 1397–1411. (doi:10.1101/gad.991602)
27. Koch U, Lehal R, Radtke F. 2013 Stem cells living with a Notch. *Development* **140**, 689–704. (doi:10.1242/dev.080614)
28. Durbin AD *et al.* 2018 Selective gene dependencies in MYCN-amplified neuroblastoma include the core transcriptional regulatory circuitry. *Nat. Genet.* **50**, 1240–1246. (doi:10.1038/s41588-018-0191-z)
29. Decaesteker B *et al.* 2018 TBX2 is a neuroblastoma core regulatory circuitry component enhancing MYCN/FOXM1 reactivation of DREAM targets. *Nat. Commun.* **9**, 4866. (doi:10.1038/s41467-018-06699-9)
30. Zeid R *et al.* 2018 Enhancer invasion shapes MYCN-dependent transcriptional amplification in neuroblastoma. *Nat. Genet.* **50**, 515–523. (doi:10.1038/s41588-018-0044-9)
31. Wang L *et al.* 2019 ASCL1 is a MYCN- and LMO1-dependent member of the adrenergic neuroblastoma core regulatory circuitry. *Nat. Commun.* **10**, 5622. (doi:10.1038/s41467-019-13515-5)
32. Peng H, Ke XX, Hu R, Yang L, Cui H, Wei Y. 2015 Essential role of GATA3 in regulation of differentiation and cell proliferation in SK-N-SH neuroblastoma cells. *Mol. Med. Rep.* **11**, 881–886. (doi:10.3892/mmr.2014.2809)
33. Hämmerle B, Yañez Y, Palanca S, Cañete A, Burks DJ, Castel V, de Mora J F. 2013 Targeting neuroblastoma stem cells with retinoic acid and proteasome inhibitor. *PLoS ONE* **8**, e76761. (doi:10.1371/journal.pone.0076761)
34. Zhang Q *et al.* 2019 Collaborative ISL1/GATA3 interaction in controlling neuroblastoma oncogenic pathways overlapping with but distinct from MYCN. *Theranostics* **9**, 86–1000. (doi:10.7150/thno.30199)
35. Soldatov R *et al.* 2019 Spatiotemporal structure of cell fate decisions in murine neural crest. *Science* **364**, eaas9536. (doi:10.1126/science.aas9536)
36. Riddick G *et al.* 2017 A core regulatory circuit in glioblastoma stem cells links MAPK activation to a transcriptional program of neural stem cell identity. *Sci. Rep.* **7**, 43605. (doi:10.1038/srep43605)
37. Bleu M *et al.* 2019 PAX8 activates metabolic genes via enhancer elements in renal cell carcinoma. *Nat. Commun.* **10**, 3739. (doi:10.1038/s41467-019-11672-1)
38. Chen Y *et al.* 2019 Bromodomain and extraterminal proteins foster the core transcriptional regulatory programs and confer vulnerability in liposarcoma. *Nat. Commun.* **10**, 1353. (doi:10.1038/s41467-019-09257-z)
39. Zhang Z *et al.* 2019 An AR-ERG transcriptional signature defined by long-range chromatin interactomes in prostate cancer cells. *Genome Res.* **29**, 223–235. (doi:10.1101/gr.230243.117)
40. Ran L *et al.* 2015 Combined inhibition of MAP kinase and KIT signaling synergistically destabilizes ETV1 and suppresses GIST tumor growth. *Cancer Discov.* **5**, 304–315. (doi:10.1158/2159-8290.CD-14-0985)
41. Ott CJ *et al.* 2018 Enhancer architecture and essential core regulatory circuitry of chronic lymphocytic leukemia. *Cancer Cell.* **34**, 982–995. (doi:10.1016/j.ccell.2018.11.001)
42. Sanda K *et al.* 2012 Core transcriptional regulatory circuit controlled by the TAL1 complex in human T cell acute lymphoblastic leukemia. *Cancer Cell.* **22**, 209–221. (doi:10.1016/j.ccr.2012.06.007)
43. Patel AP *et al.* 2014 Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401. (doi:10.1126/science.1254257)
44. Suvà ML *et al.* 2014 Reconstructing and reprogramming the tumor-propagating potential of glioblastoma stem-like cells. *Cell* **157**, 580–594. (doi:10.1016/j.cell.2014.02.030)
45. Ying M *et al.* 2011 Regulation of glioblastoma stem cells by retinoic acid: role for Notch pathway inhibition. *Oncogene* **30**, 3454–3467. (doi:10.1038/onc.2011.58)
46. Zhu X, Zhou W, Jin H, Li T. 2018 Brn2 alone is sufficient to convert astrocytes into neural progenitors and neurons. *Stem Cells Dev.* **27**, 736–744. (doi:10.1089/scd.2017.0250)
47. Graham V, Khudyakov J, Ellis P, Pevny L. 2003 SOX2 functions to maintain neural progenitor identity. *Neuron* **39**, 749–765. (doi:10.1016/S0896-6273(03)00497-5)
48. Ligon KL *et al.* 2007 Olig2-regulated lineage-restricted pathway controls replication competence in neural stem cells and malignant glioma. *Neuron* **53**, 503–517. (doi:10.1016/j.neuron.2007.01.009)
49. Godlewski J *et al.* 2008 Targeting of the Bmi-1 oncogene/stem cell renewal factor by microRNA-128 inhibits glioma proliferation and self-renewal. *Cancer Res.* **68**, 9125–9130. (doi:10.1158/0008-5472.CAN-08-2629)

50. Liu C *et al.* 2011 Wnt/beta-Catenin pathway in human glioma: expression pattern and clinical/prognostic correlations. *Clin. Exp. Med.* **11**, 105–112. (doi:10.1007/s10238-010-0110-9)
51. Zhang X, Chen T, Zhang J, Mao Q, Li S, Xiong W, Qiu Y, Xie Q, Ge J. 2012 Notch1 promotes glioma cell migration and invasion by stimulating β-catenin and NF-κB signaling via AKT activation. *Cancer Sci.* **103**, 181–190. (doi:10.1111/j.1349-7006.2011.02154.x)
52. Purow BW *et al.* 2005 Expression of Notch-1 and its ligands, Delta-like-1 and Jagged-1, is critical for glioma cell survival and proliferation. *Cancer Res.* **65**, 2353–2363. (doi:10.1158/0008-5472.CAN-04-1890)
53. Lee J *et al.* 2006 Tumor stem cells derived from glioblastomas cultured in bFGF and EGF more closely mirror the phenotype and genotype of primary tumors than do serum-cultured cell lines. *Cancer Cell* **9**, 391–403. (doi:10.1016/j.ccr.2006.03.030)
54. Qin S, Liu M, Niu W, Zhan CL. 2011 Dysregulation of Kruppel-like factor 4 during brain development leads to hydrocephalus in mice. *Proc. Natl Acad. Sci. USA* **108**, 21 117–21 121. (doi:10.1073/pnas.1112351109)
55. Carro MS *et al.* 2010 The transcriptional network for mesenchymal transformation of brain tumours. *Nature* **463**, 318–325. (doi:10.1038/nature08712)
56. Córdoba Rovira SM, Inarejos Clemente EJ. 2016 Childhood rhabdomyosarcoma. *Radiologia* **58**, 481–490. (doi:10.1016/j.rx.2016.09.003)
57. Morris JP, Yashinskie JJ, Koche R. 2019 α-Ketoglutarate links p53 to cell fate during tumour suppression. *Nature* **573**, 595–599. (doi:10.1038/s41586-019-1577-5)
58. Stewart E *et al.* 2018 Identification of therapeutic targets in rhabdomyosarcoma through integrated genomic, epigenomic, and proteomic analyses. *Cancer Cell* **34**, 411–426. (doi:10.1016/j.ccell.2018.07.012)
59. Gryder BE *et al.* 2017 PAX3-FOXO1 establishes myogenic super enhancers and confers BET bromodomain vulnerability. *Cancer Discov.* **7**, 884–899. (doi:10.1158/2159-8290.CD-16-1297)
60. Schmidt K, Glaser G, Wernig A, Wegner M, Rosorius O. 2003 Sox8 is a specific marker for muscle satellite cells and inhibits myogenesis. *J. Biol. Chem.* **278**, 29 769–29 775. (doi:10.1074/jbc.M301539200)
61. Weider M, Wegner M. 2017 SoxE factors: transcriptional regulators of neural differentiation and nervous system development. *Semin. Cell Dev. Biol.* **63**, 35–42. (doi:10.1016/j.semcdb.2016.08.013)
62. Gryder BE *et al.* 2020 Miswired enhancer logic drives a cancer of the muscle lineage. *iScience* **23**, 101103. (doi:10.1016/j.isci.2020.101103)
63. Hsieh JJ, Purdue MP, Signoretti S, Swanton C, Albiges L, Schmidinger M, Heng DY, Larkin J, Ficarra V. 2017 Renal cell carcinoma. *Nat. Rev. Dis. Prim.* **3**, 17009. (doi:10.1038/nrdp.2017.9)
64. Kaelin WG. 2007 von Hippel-Lindau Disease. *Annu. Rev. Pathol. Mech. Dis.* **2**, 145–173. (doi:10.1146/annurev.pathol.2.010506.092049)
65. Ricketts CJ *et al.* 2018 The cancer genome atlas comprehensive molecular characterization of renal cell carcinoma. *Cell Rep.* **23**, 3698. (doi:10.1016/j.celrep.2018.06.032)
66. Fletcher CDM, Bridge JA, Hogendoorn PCW, Mertens F. 2013 *WHO classification of tumours of soft tissue and bone*. Lyon, France: IARC Press.
67. Kanojia D *et al.* 2015 Genomic landscape of liposarcoma. *Oncotarget.* **6**, 42 429–42 444. (doi:10.18632/oncotarget.6464)
68. Jones RL, Fisher C, Al-Muderis O, Judson I. 2005 Differential sensitivity of liposarcoma subtypes to chemotherapy. *Eur. J. Cancer* **41**, 2853–2860. (doi:10.1016/j.ejca.2005.07.023)
69. Göransson M, Andersson MK, Forni C, Ståhlberg A, Andersson C, Olofsson A, Mantovani R, Aman P. 2009 The myxoid liposarcoma FUS-DDIT3 fusion oncoprotein deregulates nf-kappab target genes by interaction with NFKBIZ. *Oncogene.* **28**, 270–278. (doi:10.1038/onc.2008.378)
70. Knight JC, Renwick PJ, Dal Cin P, Van den Berghe H, Fletcher CD. 1995 Translocation t(12;16)(q13;p11) in myxoid liposarcoma and round cell liposarcoma: molecular and cytogenetic analysis. *Cancer Res.* **55**, 24–27.
71. Shi J, Vakoc CR. 2014 The mechanisms behind the therapeutic activity of BET bromodomain inhibition. *Mol. Cell.* **54**, 728–736. (doi:10.1016/j.molcel.2014.05.016)
72. Shen MM, Abate-Shen C. 2010 Molecular genetics of prostate cancer: new prospects for old challenges. *Genes Dev.* **24**, 1967–2000. (doi:10.1101/gad.1965810)
73. Chen Y, Sawyers CL. 2010 Coordinate transcriptional regulation by ERG and androgen receptor in fusion-positive prostate cancers. *Cancer Cell.* **17**, 415–416. (doi:10.1016/j.ccr.2010.04.022)
74. Chng KR, Chang CW, Tan SK, Yang C, Hong SZ, Sng NY, Cheung E. 2012 A transcriptional repressor co-regulatory network governing androgen response in prostate cancers. *EMBO J.* **31**, 2810–2823. (doi:10.1038/emboj.2012.112)
75. Chi P *et al.* 2010 etv1 is a lineage survival factor that cooperates with KIT in gastrointestinal stromal tumours. *Nature* **467**, 849–853. (doi:10.1038/nature09409)
76. Hirota S *et al.* 1998 Gain-of-function mutations of c-kit in human gastrointestinal stromal tumors. *Science* **279**, 577–580. (doi:10.1126/science.279.5350.577)
77. Jones DTW *et al.* 2012 Dissecting the genomic complexity underlying medulloblastoma. *Nature* **488**, 100–105. (doi:10.1038/nature11284)
78. Cho YJ *et al.* 2011 Integrative genomic analysis of medulloblastoma identifies a molecular subgroup that drives poor clinical outcome. *J. Clin. Oncol.* **29**, 424–1430. (doi:10.1200/jco.2011.29.4_suppl.424)
79. Northcott PA *et al.* 2014 Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511**, 428–434. (doi:10.1038/nature13379)
80. Hovestadt V *et al.* 2014 Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. *Nature* **510**, 537–541. (doi:10.1038/nature13268)
81. Brown JR *et al.* 2012 Integrative genomic analysis implicates gain of PIK3CA at 3q26 and MYC at 8q24 in chronic lymphocytic leukemia. *Clin. Cancer Res.* **18**, 3791–3802. (doi:10.1158/1078-0432.CCR-11-2342)
82. Landau DA *et al.* 2013 Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152**, 714–726. (doi:10.1016/j.cell.2013.01.019)
83. Cozma D *et al.* 2007 B cell activator PAX5 promotes lymphomagenesis through stimulation of B cell receptor signaling. *J. Clin. Invest.* **117**, 2602–2610. (doi:10.1172/JCI30842)
84. Hunter JE *et al.* 2016 The NF-κB subunit c-Rel regulates Bach2 tumour suppressor expression in B-cell lymphoma. *Oncogene* **35**, 3476–3484. (doi:10.1038/onc.2015.399)
85. Armstrong SA, Look AT. 2005 Molecular genetics of acute lymphoblastic leukemia. *J. Clin. Oncol.* **23**, 6306–6315. (doi:10.1200/JCO.2005.05.047)
86. Ramsay RG, Gonda TJ. 2008 MYB function in normal and cancer cells. *Nat. Rev. Cancer* **8**, 523–534. (doi:10.1038/nrc2439)
87. Blanke CD *et al.* 2008 Long-term results from a randomized phase II trial of standard- versus higher-dose imatinib mesylate for patients with unresectable or metastatic gastrointestinal stromal tumors expressing KIT. *J. Clin. Oncol.* **26**, 620–625. (doi:10.1200/JCO.2007.13.4403)
88. Dematteo RP, Heinrich MC, El-Rifai WM, Demetri G. 2002 Clinical management of gastrointestinal stromal tumors: before and after STI-571. *Hum. Pathol.* **33**, 466–477. (doi:10.1053/hupa.2002.124122)
89. Kwiatkowski N *et al.* 2014 Targeting transcription regulation in cancer with a covalent CDK7 inhibitor. *Nature* **511**, 616–620. (doi:10.1038/nature13393)
90. Ozer HG *et al.* 2018 BRD4 profiling identifies critical chronic lymphocytic leukemia oncogenic circuits and reveals sensitivity to PLX51107, a novel structurally distinct BET inhibitor. *Cancer Discov.* **8**, 458–477. (doi:10.1158/2159-8290.CD-17-0902)
91. Bai L *et al.* 2017 Targeted degradation of BET proteins in triple-negative breast cancer. *Cancer Res.* **77**, 2476–2487. (doi:10.1158/0008-5472.CAN-16-2622)
92. Baker EK *et al.* 2015 BET inhibitors induce apoptosis through a MYC independent mechanism and synergise with CDK inhibitors to kill osteosarcoma cells. *Sci. Rep.* **5**, 10120. (doi:10.1038/srep10120)
93. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, Furlong LI. 2019 The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* **48**, D845–D885.
94. Yu G, Wang LG, Yan GR, He QY. 2015 DOSE: an R/Bioconductor package for disease ontology semantic

and enrichment analysis. *Bioinformatics* **31**, 608–609. (doi:10.1093/bioinformatics/btu684)
95. Delaneau O *et al.* 2019 Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science* **364**, eaat8266. (doi:10.1126/science.aat8266)
96. Gasperini M *et al.* 2019 A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* **176**, 377–390. (doi:10.1016/j.cell.2018.11.029)
97. Grosselin K *et al.* 2019 High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nat. Genet.* **51**, 1060–1066. (doi:10.1038/s41588-019-0424-9)
98. Lieberman-Aiden E *et al.* 2009 Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293. (doi:10.1126/science.1181369)
99. Krijger P, Geeven G, Bianchi V, Hilvering C, de Laat W. 2020 4C-seq from beginning to end: a detailed protocol for sample preparation and data analysis. *Methods* **170**, 17–32. (doi:10.1016/j.ymeth.2019.07.014)
100. Li G, Sun T, Chang H, Cai L, Hong P, Zhou Q. 2019 Chromatin interaction analysis with updated ChIA-PET tool (V3). *Gene* **10**, 554. (doi:10.3390/genes10070554)
101. Raviram R, Rocha PP, Bonneau R, Skok JA. 2014 Interpreting 4C-Seq data: how far can we go? *Epigenomics* **6**, 455–457. (doi:10.2217/epi.14.47)
102. Fang R, Yu M, Li G, Chee S, Liu T, Schmitt AD, Ren B. 2016 Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Res.* **26**, 1345–1348. (doi:10.1038/cr.2016.137)
103. Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, Chang HY. 2016 HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13**, 919–922. (doi:10.1038/nmeth.3999)
104. Gryder BE, Khan J, Stanton BZ. 2020 Measurement of differential chromatin interactions with absolute quantification of architecture (AQuA-HiChIP). *Nat. Protoc.* **15**, 1209–1236. (doi:10.1038/s41596-019-0285-9)
105. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. 2015 ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **109**, 21.29.1–21.29.9. (doi:10.1002/0471142727.mb2129s109)
106. Preissl S *et al.* 2018 Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat. Neurosci.* **21**, 432–439. (doi:10.1038/s41593-018-0079-3)

Bravo-Merodio *et al. J Transl Med* (2019) 17:155
https://doi.org/10.1186/s12967-019-1912-5

Journal of Translational Medicine

RESEARCH Open Access

-Omics biomarker identification pipeline for translational medicine

Laura Bravo-Merodio[1,2], John A. Williams[1,2,3], Georgios V. Gkoutos[1,2,4,5,6,7] and Animesh Acharjee[1,2,6*]

Abstract

Background: Translational medicine (TM) is an emerging domain that aims to facilitate medical or biological advances efficiently from the scientist to the clinician. Central to the TM vision is to narrow the gap between basic science and applied science in terms of time, cost and early diagnosis of the disease state. Biomarker identification is one of the main challenges within TM. The identification of disease biomarkers from -omics data will not only help the stratification of diverse patient cohorts but will also provide early diagnostic information which could improve patient management and potentially prevent adverse outcomes. However, biomarker identification needs to be robust and reproducible. Hence a robust unbiased computational framework that can help clinicians identify those biomarkers is necessary.

Methods: We developed a pipeline (workflow) that includes two different supervised classification techniques based on regularization methods to identify biomarkers from -omics or other high dimension clinical datasets. The pipeline includes several important steps such as quality control and stability of selected biomarkers. The process takes input files (outcome and independent variables or -omics data) and pre-processes (normalization, missing values) them. After a random division of samples into training and test sets, Least Absolute Shrinkage and Selection Operator and Elastic Net feature selection methods are applied to identify the most important features representing potential biomarker candidates. The penalization parameters are optimised using 10-fold cross validation and the process undergoes 100 iterations and a combinatorial analysis to select the best performing multivariate model. An empirical unbiased assessment of their quality as biomarkers for clinical use is performed through a Receiver Operating Characteristic curve and its Area Under the Curve analysis on both permuted and real data for 1000 different randomized training and test sets. We validated this pipeline against previously published biomarkers.

Results: We applied this pipeline to three different datasets with previously published biomarkers: lipidomics data by Acharjee et al. (Metabolomics 13:25, 2017) and transcriptomics data by Rajamani and Bhasin (Genome Med 8:38, 2016) and Mills et al. (Blood 114:1063–1072, 2009). Our results demonstrate that our method was able to identify both previously published biomarkers as well as new variables that add value to the published results.

Conclusions: We developed a robust pipeline to identify clinically relevant biomarkers that can be applied to different -omics datasets. Such identification reveals potentially novel drug targets and can be used as a part of a machine-learning based patient stratification framework in the translational medicine settings.

Keywords: Biomarker, -Omics, Regularization, Feature selection, Translational medicine

*Correspondence: a.acharjee@bham.ac.uk
[1] College of Medical and Dental Sciences, Institute of Cancer and Genomic Sciences, Centre for Computational Biology, University of Birmingham, Birmingham B15 2TT, UK
Full list of author information is available at the end of the article

Background

Translational medicine (TM) [1–3] is an emerging and fast growing area of research that aims to facilitate medical or biological advances efficiently from the scientist to the clinician. TM approaches uses diagnostic tools and treatments by commonly employing interdisciplinary frameworks, in a highly collaborative manner to reach out to the patient community for disease treatment, stratification and prevention. A notion central to TM is to narrow the gap between basic science and applied science in terms of time, cost and early diagnosis of the disease state. Over the last few decades the influx of untargeted -omics (phenomics, transcriptomics, metabolomics, epigenomics, lipidomics and others) datasets have enabled the identification of biological markers of disease (so-called biomarkers) [4], and have become one of the main avenues towards discovery within TM. The identification of the disease biomarkers from -omics data does not only facilitate the stratification of patient cohorts but also provides early diagnostic information to improve patient management and prevent adverse outcomes. However, biomarker identification, a task that is commonly comprised of biological and computational processes, needs to be robust and reproducible if it is to be clinically useful and actionable in patient-care settings or in response to new therapies. Therefore, a robust unbiased computational framework is necessary to identify biological signals that can reveal potential novel biomarkers.

In -omics literature, there has been a recent trend towards the identification of data pre- and post-processing steps. For example, Satagopam et al. [5] developed an infrastructure comprised by a combination of web services, tranSMART, Galaxy, and MINERVA platforms. Narayanasam et al. [6] developed an integrated reference-independent analysis of metagenomic and metatranscriptomic data for the analysis of microbiome derived datasets. Feng [7] developed a proteomics pipeline called Firmiana. Firmiana is a cloud platform that allows scientists to deposit mass spectrometry (MS) raw files online and performs automated bioinformatic analyses on the uploaded data. Such existing robust pipelines for analysing -omics data are often either focused on specific -omics data or can be used only for either classification or regression purposes. For example, Xia et al. [8] developed a workflow for quantitative metabolomics datasets, Acharjee et al. [9] developed an -omics fusion tool but focused on metabolomics data in regression mode only. Hermida et al. [10] developed a pipeline based on transcriptomics data called Confero that extracts gene lists from research papers and performs automatic extraction and storage of gene sets. While this is useful for downstream analysis, there is a need to combine these approaches and deal with multiple types of outcome data as well as consider their categorical or continuous nature. In some cases, the complexity of machine learning models associated visualizations used hinder the interpretability of the results and therefore impair their translation into clinical science.

In this study, we develop a pipeline that includes two machine learning algorithms, inspired by simple linear models, coupled with follow-up approaches for systematic data analysis. Our systematic analysis includes data quality checks, identification of important features, as well as combinatorial and stability analyses. We applied and validated our pipeline with three different previously published -omics datasets. Our approach successfully identified the markers reported in the literature as well as potential novel markers.

Materials and methods

Our pipeline is composed of statistical machine learning modules whose methods are described below. Additionally, we applied and validated our pipeline against three independent published datasets; two RNA microarray datasets and one lipidomics experiment.

Machine learning methods

We used two feature selection methods, LASSO [11] and Elastic Net [12]. These are two forms of regularization methods that are able to automatically select the features from the dataset and hence provide a sparse solution. Regularization works in the following way: Starting from simple linear regression models we consider $x_1 \ldots x_p$ as x number of predictor variables (features) and y as an outcome or response variable:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots \hat{\beta}_p x_p \tag{1}$$

Here, the outcome of model fitting produces the vector of estimated regression coefficients through ordinary least squares (OLS), with the objective function as the minimum of the residual sum of the squares (RSS) equation (Eq. 2). The values minimizing the function are the estimated regression coefficients (β).

$$Residualsumofsquares(RSS) = \sum_{i=1}^{N} \left(y_i - x_i^T \beta\right)^2 \tag{2}$$

In regularization methods, an extra term is added (Eq. 3) and so the new objective function to minimize becomes:

$$RSS(\beta) + p\lambda(\beta) \tag{3}$$

Here p is a function to penalize and λ forms the penalty/regularization parameter. The penalty function λ controls the trade-off between likelihood and penalty and so influences the variables to be selected. The higher the value of λ, the fewer number of variables are selected and vice versa. The differences between regularization methods

lie in the different functions p they penalize. In LASSO, the penalty is applied to the sum of the absolute values of the regression coefficients (L1 norm). Mathematically, we can write this as:

$$\underset{\beta \in R^p}{minimize} \frac{1}{2}\left\|y - X\beta\right\|_2^2 + \lambda\|\beta\|_1 \quad (4)$$

The left part of the equation is the normal least squares criterion, whereas the right part is the penalized sum of the absolute values of the regression coefficients.

In Ridge regression [13], the precursor of LASSO, the penalization p is incurred in the L2 norm of the coefficients (sum of the squares). In this case, selection is not sparse since coefficients are never zero but close and so, a rank of features based on the penalised regression coefficients, is produced. Elastic Net [12], on the other hand, is a mixed version of both LASSO and Ridge (Eq. 5). It allows for the sparse representation, similarly to LASSO, and theoretically improves its performance in $p \gg n$ cases with high collinear groups of features by allowing grouped selection or de selection of correlated variables. LASSO instead tends to select only one "random" variable from the group of pairwise correlations. EN is created through the merging of both Ridge and LASSO penalizations (Eq. 5). A different representation of the same equation can be seen below (Eq. 6), with a single parameter α regulating the relationship between Ridge and LASSO. When α is equal or closer to 0 we have a stronger penalization and so a solution closer or equal to LASSO whereas, when α is equal or closer to 1, the behaviour resembles Ridge.

$$\underset{\beta \in R^p}{minimize} \frac{1}{2}\left\|y - X\beta\right\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2 \quad (5)$$

$$\underset{\beta \in R^p}{minimize} \frac{1}{2}\left\|y - X\beta\right\|_2^2 \; subject \; (1-\alpha)\|\beta\|_1 + \alpha\|\beta\|^2 \quad (6)$$

This combination of LASSO and EN methods comprise the backbone of our pipeline and the construction is described below.

Pipeline construction and follow-up analysis methods

All analyses were performed in the R statistical computing (R version 3.4.3) environment [14]. All R packages can be found in our project's github repository stated below. The necessary software dependencies are described in the README file located in the repository. All analyses can be performed on a standard PC environment with the run time increasing with larger datasets. For example, an analysed RNA microarray acute myeloid leukaemia (AML) dataset described below took 8 h to complete, but at no time did the R environment use more than 1 Gb of RAM. The workflow is embedded in a R Markdown file which, when altered with a user's working and output directories and the name of the input data file, runs the analysis in real time. After running, it compiles a PDF report of containing both all code generated and figures produced. Figures generated include analogues to Figs. 2 and 3. Additionally, ROC AUC curves are generated via stability analyses for individual selected features as well as combinations shown to be significant in predicting binary outcome. Importantly, a list of significant features (genes, metabolites, etc.) is printed in the PDF report. These lists can easily be copied so as to be used as input for pathway and ontology enrichment analyses. For the purposes of our validation studies, pathway analysis and ontology enrichment wereperformed with the EnrichR tool with default settings (analysed on Sept 7, 2018) [15].

Lipidomics data

To assess the performance of our framework, we employed a previously published lipidomics dataset from Acharjee et al. [16]. The lipidomics data was generated from The Cambridge Baby Growth Study (CBGS), a prospective observational birth cohort. For details about the processes related to the data generation as well as sample information please check Acharjee et al., and Prentice et al. [16, 17].

From the CBGS cohort we used 3 different datasets, namely CBGS-1, CBGS-2 and POPS (Pregnancy Outcome Prediction Study). All data was obtained from dried blood spots and generated with direct infusion high-resolution mass spectrometry (HRMS).

A summary of the cohort is listed below (Table 1).

Transcriptomic data

A pancreatic ductal adenocarcinoma (PDAC) microarray expression dataset (n = 36 control, n = 36 cases) GSE15471, [18], as well as microarray expression data from a three-cohort study of acute myeloid leukaemia (AML) cell lines with n = 404 AML samples and n = 138 control samples, excluding a third transitional cohort

Table 1 Cohort statistics of samples analysed form the Cambridge Baby Growth Study (CBGS) and Pregnancy Outcome Prediction Study (POPS)

Cohort name	Sample information			Total sample number (n)
	Formula milk (FM)	Breast milk(HM)	Mixed (FM and HM)	
CBGS-1	85	87	67	239
CBGS-2	43	25	27	95
POPS	16	11	13	40

of MDS samples GSE15061 [19] was analysed. In each case, the Robust Multichip Average standardized Affymetrix Human Genome U133 Plus 2.0 data submitted by the authors to the Gene Expression Omnibus was taken as inputalong with class information indicating case or control condition. Pre-processing of microarray data included; (a) taking the median of duplicate probes across all conditions to yield one unique probe per experiment, (b) collapsing rows of probes belonging to identical genes and taking maximum expressed probe via the WGCNA R package [20], and (c) testing features (samples) for low variance via the caret R package and removing those with near-zero variance among all genes [21]. The resulting numerical matrix of normalized gene expression values was used as input for each experiment, yielding 22,880 genes and 542 samples for the AML dataset, and an equal number of genes and 78 samples for the pancreatic cancer dataset. Normalized data matrices produced both for validating the reproducibility of our pipeline, as well as the ones used as example input, are available in our project github repository. The different steps comprising our pipeline are described in the results section.

Availability of code

All code and functions are available on our hosted GitHub repository: https://github.com/jaw-bioinf/Biomarker_Identification/.

Results

Pipeline features

A graphical depiction of our workflow can be seen in Fig. 1. Our pipeline can be divided in the three modules that we describe below with the purple data quality module offering different options depending on the type of data introduced e.g. microarray and generic/other data.

Data quality module

The data quality module consists of different checks on both features and samples to exclude or data based on the amount of missingness, and to standardize data to make measurements of features in different experiments comparable. Missing value imputation and normalisation steps can be implemented as needed by end users. Normalization methods can be bypassed if RNA microarray datasets are downloaded pre-processed directly from repositories. For these and other microarray datasets, optional filtering steps to reduce dimensionality include many-to-one probe-to-gene mapping. Features may be further reduced by using external tools or expert biological insight to exclude features, but such reduction is optional. In all cases, the input for the 'feature selection' segment of the pipeline is a matrix (file or data frame) of features (potential biomarkers) as well as a set of samples along with a target or outcome variable.

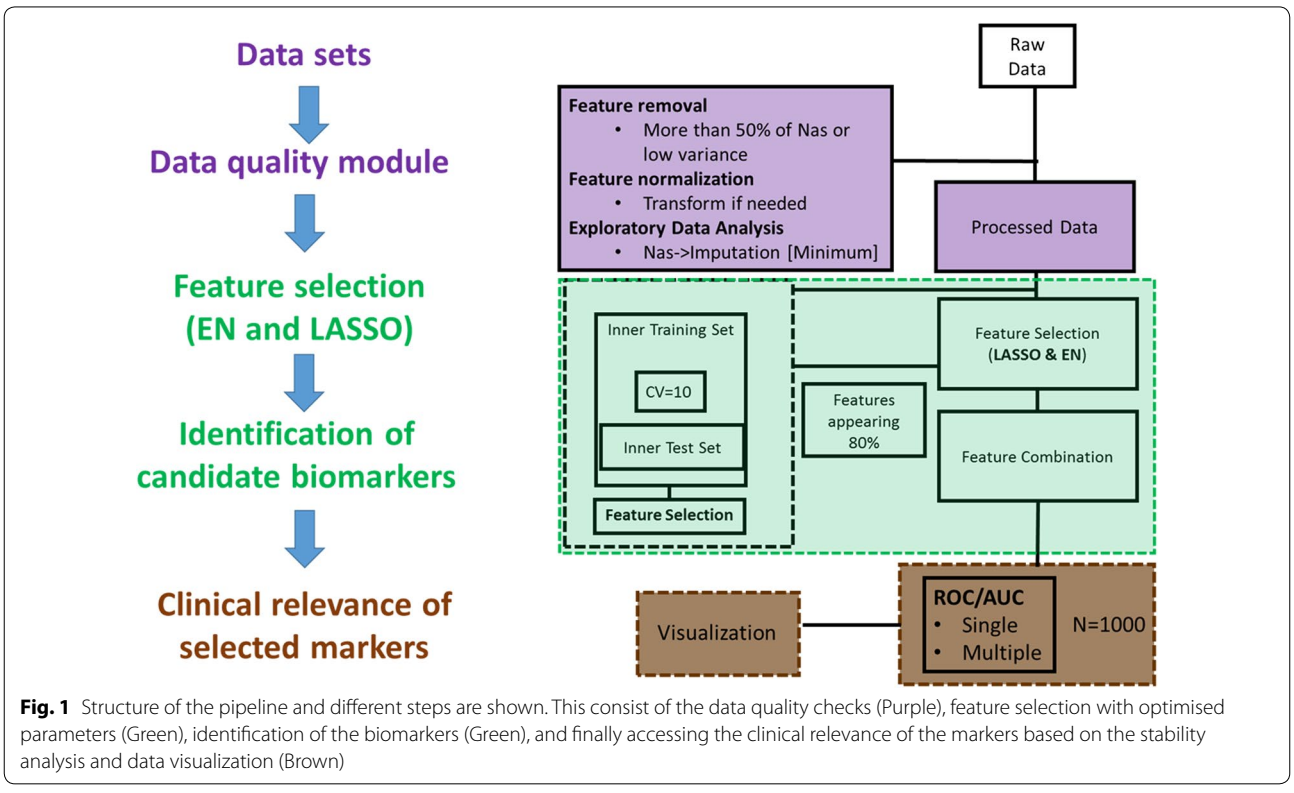


Fig. 1 Structure of the pipeline and different steps are shown. This consist of the data quality checks (Purple), feature selection with optimised parameters (Green), identification of the biomarkers (Green), and finally accessing the clinical relevance of the markers based on the stability analysis and data visualization (Brown)

Feature selection and parameter optimization

In order to apply the LASSO and EN algorithms for biologically relevant feature selection, we need to optimize the penalty parameter associated with each of the methods in an unbiased way. To achieve this, the pipeline divides samples randomly into a training set composed of 75% of the total number of the samples and a test set consisting of the remaining 25% samples. To ensure optimal training in real-world datasets, all data splits retained class balances of the target variable, so each split reflected a proportion of the target observed in the underlying dataset. We note that class balancing measures, such as boosting or under sampling, are not used to artificially balance training/testing data in each split (outer loop set). Then we apply a 10-fold cross validation on the training set (inner loop set) aiming to have an optimised penalty parameter that can be plugged into the LASSO and EN models. Mathematically, LASSO and EN models can be defined by using a single penalty function "α" [22] (Eq. 6). For example, by using a penalty parameter $\alpha = 1$, we are applying the LASSO algorithm, whereas $\alpha = 0.5$ will perform Elastic Net. A high value for the penalty parameter (α) will result in a strong penalty and hence fewer variables will be selected.

Identification of candidate biomarkers

Our pipeline iterates model creation 100 times and selects the features that appear more than 90 times in the analysis, as these we deem to be the more significant for the classification model. Moreover, in order to better understand the relationship between the features selected and the outcome variable analysed, a display of the weight (β coefficients) distribution per model (see Additional file 1) and a box plot of the class differences per feature is generated (Fig. 3b). These selected features are then considered potential candidate biomarkers. In order to ascertain their validity as biomarkers, their performance is evaluated both alone and in combination.

Performance evaluation and visualization

In order to investigate the performance of the selected markers, our pipeline performs stability analysis through a permutation test. This consists of the randomization of the label features, resulting in incorrect sample labels for predictions and generating models with ROC AUC values showing a performance subject to the random distribution. Both the real model and permutation tests are produced by sampling 1000 random training and test sets, then using simple machine learning models to consider the fit of data, with ROC AUC performance results plotted as density plots alongside their means and standard deviation. The ROC AUC offers a graphical overview of the diagnostic ability of binary classifiers with varying thresholds. In addition to this, more information on the predictive ability of the model is obtained through the calculation of the sensitivity, specificity, precision and accuracy values.

Validation of the approach

Lipidomics data

We applied our pipeline in the published lipidomics data available from three cohorts: the Cambridge Baby Growth Study (CBGS1and CBGS2) and the Pregnancy Outcome Prediction Study (POPS). Our objective was to identify potentially nutritional lipid biomarkers for the classification of babies fed with Formula, Human or a mix of Human and Formula milk. In Fig. 2a, we display the frequency of appearance of the lipids in 100 different Elastic Net models of classification between Formula and Human milk nutrition from CBGS2 data. Figure 2b shows the same results but for LASSO. It can be seen that EN allows for a less stringent solution with more features appearing. Additional file 2: Table 2 reveals the high-ranking lipids identified by our approach as well as their associated nutritional outcomes.

For those selected features, performance evaluation was then performed. Results can be seen in Fig. 3a where, a combination of the three selected lipids SM(39:1), SM(32.1) and SM(36.2) shows a significant improvement in the models ability to classify between Human milk and Formula and Human mixed milk (from a 0.5 in permuted data to a 0.83 ROC AUC value) in the CBGS1 data. Moreover, as seen in Fig. 3b direct visualization of the selected lipids in a box plot, allows for a clear display of the differing prevalence of this feature in babies fed with these different milk nutrition and so explaining its selection and inclusion in the classification model. These plots are easy to interpret and hence reach out to a non-expert domain. Moreover, our analysis revealed a consistent biomarker robustness, between HM and FM diets, across three different cohorts, summarised in the Additional file 2. For example, SM(39:1) is identified as a robust biomarker for segregating infants on HM vs. FM diets (Additional file 2: Table 2).

Transcriptomics data

In the pancreatic cancer dataset GSE15471, top features selected included the following 20 genes: SULF1, COL10A1, MIR34AHG, INHBA, COL8A1, FN1, THBS2, NOX4, NTM, RASAL2, ADAMTS12, CAPG, CTHRC1, FAP, VCAN, SLPI, WISP1, LTBP1, GPRC5A, TIMP1. Our biological pathway analysis revealed a class of biomarkers enriched in several known cancer pathways. After stringent multiple testing correction, the following pathways were identified as being enriched:

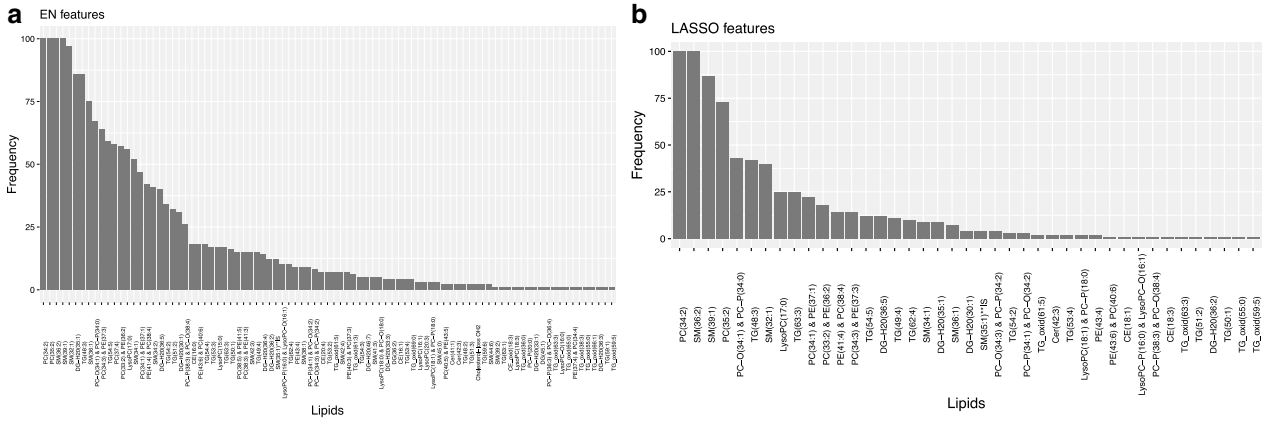


Fig. 2 **a** The frequency of the lipids (y axis) ranked by their selection out of 100 randomized sampling splits using Elastic Net. **b** A similar analysis is shown using LASSO. Features identified in both analyses are used for further investigation

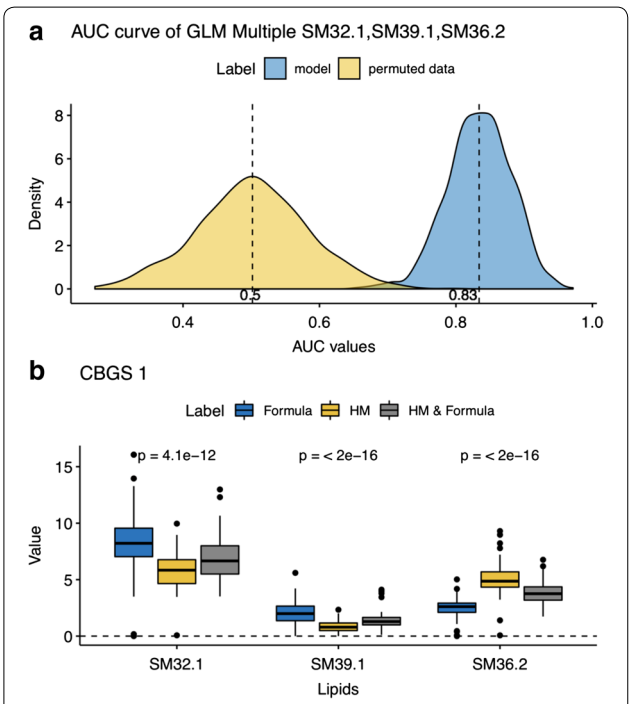


Fig. 3 **a** Stability of combination of lipids SM(39:1), SM(32:1) and SM(36:2) (blue) versus random sampling-derived null distribution (yellow) in CBGS1, formula milk versus formula milk and human milk model prediction. X-axis is the area under the receiver operating characteristic curve (ROC), and dashed lines represent the mean over all 1000 sampled trials conducted. **b** Probability values showing a significant difference between the measured levels of selected lipids between all the nutritional classes (formula vs. human milk vs. mixed)

- Senescence and Autophagy in Cancer
- Integrated Pancreatic Cancer Pathway
- TGF Beta Signalling Pathway (mouse)
- miRNA targets in ECM and membrane receptors
- TGF Beta Receptor Signalling (human).

The gene set was further enriched in the Human Proteome Map in adult Esophagus, Lung, and Pancreas tissues, indicating a potential cross-talk among tissue-specific cancer pathways. The full results of our pipeline, including feature ranking graphs and ROC AUC, sensitivity, specificity, and accuracy scores of each variable, as well as all the executed code are included in Additional file 1.

In addition to the validation of our method against pancreatic cancer, a validation was performed against an acute myeloid leukaemia (AML) cohort. This analysis revealed sixty genes contributing significantly to AML (see Additional file 3). This gene set interacts with several AML-associated transcription factors, including NKX2-3, HOXA7, and MYB. The analysis of genes active in cell lines available in the Cancer Cell Line Encyclopedia [23] revealed that our derived gene set is significantly enriched in multiple haematopoietic and lymphoid tissue lines. Additionally, investigation of the presence of our depicted genes in biological pathways, annotated in the KEGG database [24], confirmed their known AML-gene associations mediated by the 'Hematopoietic stem cell lineage' and 'Transcriptional misregulation in cancer' pathways. Further results from these analyses are presented in Additional file 3.

Discussion

We developed a systematic way of analysing -omics datasets to identify potential biomarkers from large-scale -omics datasets. We used three different datasets (two transcriptomics and one lipidomics) to validate our approach by identifying potential markers or signatures

and comparing with existing markers found in the literature.

Lipidomics data

Acharjee et al. [16] investigated and identified relevant lipid biomarkers that were able to predict infant feeding patterns. A narrow down list of candidate biomarkers was produced based on a combination of supervised Random Forests and iterative backward elimination. In our analysis, we used two methods, LASSO and EN, that perform automatic feature selection. Our analysis revealed two types of relevant lipids; four Sphingomyelin and two Phosphocholine. Out of these six, three were previously reported by Acharjee et al. [16]. For our validation analysis, selected features were singled out in one step, whereas in the previous study a two-step procedure was employed. Moreover, to assess the stability of the relevant identified features, we employed multiple sampling and permutation testing to test against an empirical null distribution based on ROC AUC scores (Fig. 3a).

Transcriptomics data

In order to assess the wider applicability of our approach for identifying target molecules in different types of -omics data, we also applied it in two transcriptomic datasets, one for Acute Myeloid Leukemia and one for Pancreatic Ductal AdenoCarcinoma (PDAC). Whereas the lipidomic analysis could be validated by expert curation, our transcriptomic analyses were validated via external pathway and ontology gene set enrichment tools.

AML driver gene analysis revealed a set of genes known to be enriched for targets of the MYB transcription factor. MYB is known to play a crucial role in hematopoietic stem cell cycles, including proliferation and survival, and recent research has shown that AML-specific microRNAs target c-MYB [25]. Additionally, a potentially drugable compound targeting MYB was recently discovered [26], highlighting the clinical role of MYB targets. By highlighting the genes which are both predictors of AML and enriched as a set for MYB targeting, we have identified a set of novel gene targets of the MYB transcription factor.

The genes identified by our pipeline are often discussed in pancreatic cancer literature [27]. Not only did we identify gene sets in relevant tissues which are, in combination, highly discriminative between pancreatic cancer and control, but the in-built multivariate analysis revealed interacting networks which model differences between cancer and control patient data better than single genes alone. Our analysis also highlighted the crosstalk between autophagy and certain cancer types. Given the prevalence of autophagy pathways perturbed in pancreatic cancers, this result confirms recent novel studies demonstrating autophagic control of pancreatic cancer metabolism [28, 29].

Workflow features

In high-dimensional -omics data analysis, we are interested in finding a relevant smaller subset of variables that are associated with the response (a clinical phenotype). Procedures to identify such smaller subsets are called variable or feature selection procedures. By employing such procedures, it is possible to reduce the dimensionality of the data [30]. Moreover, feature selection can assist in removing noise variables (variables which have no predictive power for the response variable) in the dataset. More specifically, typical reasons to employ feature selection procedure include: large number parameters, features or variables (p) compared to the number of the samples or individuals (n) and correlated features.

One advantage of using feature selection algorithms is that the final model is built automatically, including only those biomarkers which are useful in predicting patient condition. Thus, we do not have to rely on the cut off for selection of genes and metabolites upfront. All the estimates are decided based on either biomarkers' effects. However, one of the drawbacks of this method lies with the selection of the appropriate penalty parameters. Failure to decide on appropriate penalty factor will result in underfitting or over fitting of the results. To address this, we split the data into two subsets, training and testing. Within the training subset, we estimated the penalty factor by using ten-fold cross validation. The optimized model was then fit to the unseen testing subset.

While either LASSO or EN can be used for both classification and regression tasks, our method focused on validation tasks based on binary outcomes (classification). In a planned update of the software accompanying our method, we will enable users to switch between classification and regression tasks. Users will also be able to choose between different feature selection algorithms and machine learning models including Random Forests, Artificial Neural Networks, and Deep Leaning which can capture alternative patterns of interactions in the data that we might miss out with regularized linear models [31]. We also plan to implement our code in a portable Docker environment to eliminate the need for end users from dealing with version control and software dependencies. Lastly, it should be noted that our model currently accepts numerical variables whilst categorical variables should be dummy (one-hot) encoded.

A unique strength of our approach lies with the provision of automated pre-processing and feature selection. Based on our approach, we were able to reduce the number of potential causative genes in each experiment

to under 100 (from an input of over 22,000 genes) whilst the high confidence selections were reduced to less than 15. This robust selection creates a useful feature for end users, eliminating the need to pre-filter data based on perceived biological knowledge thus eliminating bias.

Future trends

Multi -omics data integration

To completely understand the underlying biological mechanisms driving diverse phenotypes, a multi-omics approach is often necessary. However, this is a challenging step due to the data size, measurements, and data analysis involved. Different approaches are currently suggested in the literature to link or integrate them. For example, Shen et al. used multi-omics datasets which include copy number, gene expression, and methylation data from TCGA in an unsupervised matrix factorization algorithms using the software i-Cluster [32]. Seoane et al. used a pathway-based data integration framework for prediction of breast cancer progression. They used multiple kernel learning supervised learning methods on multi-omics datasets that includes clinical data, gene expression and copy number data [33]. A similar method was further applied by Zhu et al. to integrate somatic mutation, DNA copy number, DNA methylation, mRNA and miRNA expression datasets from TCGA [34]. Acharjee et al. used Random Forests to integrate clinical, lipidomics, and metabolomics datasets. They first selected features from metabolomics and lipidomicsdataset and linked selected features by correlation analysis [35]. Pedersen et al. developed a computational framework to integrate multi-omics datasets that included human phenotype, serum metabolome and gut microbiome data. This framework allowed for a stepwise flexible choice of methods, adaptable to different -omics datasets with feature selection as one of the important first step. Additional examples of multiomics integration include linking genome, metabolome and gut microbiome [36], and the linking of somatic mutations, RNA expression, DNA methylation and ex vivo drug responses [37].

In addition to -omics datasets, there are other unstructured clinical phenotypic datasets such as medical images, electronic health records, and medical questionnaires. These pose new challenges for data integration and reproducibility that needs standardization and to put into clinical practices [38]. Proposed strategies for integrating these data into our current pipeline include deriving numerical features from these unstructured data, for example by creating vectors of word representations with word2vec models [39].

Single cell sequencing

It is worth mentioning that certain areas of precision medicine benefit greatly from incorporating single-cell sequencing data, especially cancer. While multiple -omics approaches can be used with single-cell sequencing [40, 41], RNA-Sequencing applied to single-cell data has been used extensively, and we will focus our discussion on this area [42–44]. Single-cell transcriptomics (scRNA-Seq) have potential for monitoring patient response to treatment and characterizing lineage-specific mutations which may respond to variable treatment protocols. Before incorporating scRNA-Seq data into the pipelines described above, experimentalists and analysists must be aware of several differences in protocol which affect normalization of scRNA-Seq data. Stegele et al. [45] produced a fundamental review of challenges which are currently being addressed by the community. In essence, data must be carefully curated, select data must be normalized after additional quality controls not applicable to bulk RNA-Seq. This may necessitate the inclusion of synthetic or alternate species controls (spike-ins) in the sequencing experiments not always used in bulk data analysis. After normalization, populations of cells may be identified by several unsupervised learning methods, from clustering to tSNE [46]. Our pipeline may add value to single-cell analyses by picking up at this stage and integrating count matrices of cell-types separated by clustering approaches. With the separation of cell populations, dominated by driver genes characterizing cell types or disease states as labels, our pipeline can be applied to select gene transcripts which act as biomarkers. With these biomarkers identified, subsequent patient monitoring may be applied to surveil tissues for tumour progression and guide the application of treatment or help reveal mechanisms in cells which survive treatment after resequencing [44]. Finally, it is worth mentioning too, that useful clinical translation will only follow from a better understanding of the underlying biological mechanisms for the biomarker´s discovered.

The application of single-cell omics to our pipeline may be useful in model organism and basic research to guide future translational projects or prioritize experiments for biomarker validation.

Conclusion

We present a data-driven, generalizable, robust, low-bias machine learning workflow that generates easily interpretable outputs and focus on simple visualizations aiming at actionable biomarker discovery. We believe that our workflow will help researchers to identify significant explanatory features of experimental -omics data, reducing the search space for good candidates for experimental

validation and follow up. Robustly optimizing feature selection to changes in data perturbation provides a high confidence in the selection of potential novel features, which forms a crucial advantage in translational medicine applications.

Additional files

Additional file 1. R Markdown analysis results from the workflow developed on GSE15471.

Additional file 2. Lipids identified in three cohorts are listed with different category. Category A: HM vs. Mixed (FM and HM combined) feeding; Category B: FM vs. Mixed (FM and HM combined) feeding; Category C: HM vs. FM.

Additional file 3. R Markdown analysis results from the workflow developed on GSE15061.

Abbreviations

AML: acute myeloid leukemia; CV: cross validation; EN: Elastic Net; FM: formula milk; HM: human milk; miRNAs: microRNAs; LASSO: Least Absolute Shrinkage and Selection Operator; PDAC: pancreatic ductal adenocarcinoma; ROC AUC : Receiver Operating Characteristic Area Under the Curve; RSS: residual sum of squares; SM: sphingomyelin; SNP: single nucleotide polymorphism; PC: phosphatidylcholine; TM: translational medicine.

Acknowledgements

The authors would like to thank their fellow group members for advice throughout this project.

Authors' contributions

AA conceived and designed the data analysis strategy; JAW/LBM acquired materials and performed data analysis; AA and GVG supervised the study; all authors co-wrote, edited and reviewed the manuscript. All authors read and approved the final manuscript.

Funding

This study was supported by National Institute for Health Research (NIHR) Surgical Reconstruction and Microbiology Research Centre (SRMRC), Birmingham. GVG also acknowledges support from H2020-EINFRA (731075) and the National Science Foundation (IOS:1340112) as well as support from the NIHR Birmingham ECMC, the NIHR Birmingham Biomedical Research Centre and the MRC HDR UK. JAW also acknowledges support from the National Human Genome Research Institute of the National Institutes of Health under Award Number UM1HG006370. LBM is funded by the Wellcome Trust 4-year studentship program in mechanisms of inflammatory disease (MIDAS; 108871). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the Medical Research Council or the Department of Health.

Availability of data and materials

The datasets analysed during the current study are available in the Gene Expression Omnibus. Transcriptomic datasets accessed include https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15471 and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15061. Lipidomics data were analysed from the supplementary materials of following article: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5272886/.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] College of Medical and Dental Sciences, Institute of Cancer and Genomic Sciences, Centre for Computational Biology, University of Birmingham, Birmingham B15 2TT, UK. [2] Institute of Translational Medicine, University Hospitals Birmingham NHS Foundation Trust, Birmingham B15 2TT, UK. [3] Mammalian Genetics Unit, Medical Research Council Harwell Institute, Harwell Campus, Didcot OX11 0RD, UK. [4] MRC Health Data Research UK (HDR UK), London, UK. [5] NIHR Experimental Cancer Medicine Centre, Birmingham B15 2TT, UK. [6] NIHR Surgical Reconstruction and Microbiology Research Centre, Birmingham B15 2TT, UK. [7] NIHR Biomedical Research Centre, Birmingham B15 2TT, UK.

Received: 6 October 2018 Accepted: 8 May 2019
Published online: 14 May 2019

References

1. Howells DW, Sena ES, Macleod MR. Bringing rigour to translational medicine. Nat Rev Neurol. 2014;10:37–43.
2. Han H. Diagnostic biases in translational bioinformatics. BMC Med Genomics. 2015;8:46.
3. Fang FC, Casadevall A. Lost in translation—basic science in the era of translational research. Infect Immun. 2010;78:563–6.
4. Mischak H, Allmaier G, Apweiler R, Attwood T, Baumann M, Benigni A, et al. Recommendations for biomarker identification and qualification in clinical proteomics. Sci Transl Med. 2010;2:46ps42.
5. Satagopam V, Gu W, Eifes S, Gawron P, Ostaszewski M, Gebel S, et al. Integration and visualization of translational medicine data for better understanding of human diseases. Big Data. 2016;4:97–108.
6. Narayanasamy S, Jarosz Y, Muller EEL, Heintz-Buschart A, Herold M, Kaysen A, et al. IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. Genome Biol. 2016;17:260.
7. Feng J, Ding C, Qiu N, Ni X, Zhan D, Liu W, et al. Firmiana: towards a one-stop proteomic cloud platform for data processing and analysis. Nat Biotechnol. 2017;35:409–12.
8. Xia J, Wishart DS. Using MetaboAnalyst 3.0 for comprehensive metabolomics data analysis. Curr Protoc Bioinform. 2016;55:14.10.1–10.91.
9. Acharjee A, Finkers R, Visser RG, Maliepaard C. Comparison of regularized regression methods for ~ omics data. Metabolomics. 2013;3:1–9.
10. Hermida L, Poussin C, Stadler MB, Gubian S, Sewer A, Gaidatzis D, et al. Confero: an integrated contrast data and gene set platform for computational analysis and biological interpretation of omics data. BMC Genomics. 2013;14:514.
11. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B Methodol. 1996;58:267–88.
12. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005;67:301–20.
13. Hoerl AE. Application of ridge analysis to regression problems. Chem Eng Prog. 1962;58:54–9.
14. R Core Team. R: a language and environment for statistical computing. [Internet]. Vienna: R Foundation for Statistical Computing; 2013. http://www.R-project.org/.
15. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinform. 2013;14:128.
16. Acharjee A, Prentice P, Acerini C, Smith J, Hughes IA, Ong K, et al. The translation of lipid profiles to nutritional biomarkers in the study of infant metabolism. Metabolomics. 2017;13:25.
17. Prentice P, Koulman A, Matthews L, Acerini CL, Ong KK, Dunger DB. Lipidomic analyses, breast- and formula-feeding, and growth in infants. J Pediatr. 2015;166(276–281):e6.
18. Rajamani D, Bhasin MK. Identification of key regulators of pancreatic cancer progression through multidimensional systems-level analysis. Genome Med. 2016;8:38.
19. Mills KI, Kohlmann A, Williams PM, Wieczorek L, Liu W, Li R, et al. Microarray-based classifiers and prognosis models identify subgroups with distinct clinical outcomes and high risk of AML transformation of myelodysplastic syndrome. Blood. 2009;114:1063–72.
20. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinform. 2008;9:559.

21. Wing MKC from J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, et al. caret: Classification and Regression Training [Internet]; 2018. https://CRAN.R-project.org/package=caret.
22. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33:1–22.
23. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483:603–7.
24. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.
25. Hornick NI, Doron B, Abdelhamed S, Huan J, Harrington CA, Shen R, et al. AML suppresses hematopoiesis by releasing exosomes that contain microRNAs targeting c-MYB. Sci Signal. 2016;9:ra88.
26. Uttarkar S, Frampton J, Klempnauer K-H. Targeting the transcription factor Myb by small-molecule inhibitors. Exp Hematol. 2017;47:31–5.
27. Ma X, Liu Y, Liu Y, Alexandrov LB, Edmonson MN, Gawad C, et al. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. Nature. 2018;555:371–6.
28. Perera RM, Stoykova S, Nicolay BN, Ross KN, Fitamant J, Boukhali M, et al. Transcriptional control of autophagy-lysosome function drives pancreatic cancer metabolism. Nature. 2015;524:361–5.
29. Yang M-C, Wang H-C, Hou Y-C, Tung H-L, Chiu T-J, Shan Y-S. Blockade of autophagy reduces pancreatic cancer stem cell activity and potentiates the tumoricidal effect of gemcitabine. Mol Cancer. 2015;14:179.
30. Clarke R, Ressom HW, Wang A, Xuan J, Liu MC, Gehan EA, et al. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nat Rev Cancer. 2008;8:37–49.
31. Kong Y, Yu T. A deep neural network model using random forest to extract feature representation for gene expression data classification. Sci Rep. 2018;8:16477.
32. Shen R, Mo Q, Schultz N, Seshan VE, Olshen AB, Huse J, et al. Integrative subtype discovery in glioblastoma using iCluster. PLoS ONE. 2012;7:e35236.
33. Seoane JA, Day INM, Gaunt TR, Campbell C. A pathway-based data integration framework for prediction of disease progression. Bioinform Oxf Engl. 2014;30:838–45.
34. Zhu B, Song N, Shen R, Arora A, Machiela MJ, Song L, et al. Integrating clinical and multiple omics data for prognostic assessment across human cancers. Sci Rep. 2017;7:16954.
35. Acharjee A, Ament Z, West JA, Stanley E, Griffin JL. Integration of metabolomics, lipidomics and clinical data using a machine learning method. BMC Bioinform. 2016;17:37–49.
36. Bakker OB, Aguirre-Gamboa R, Sanna S, Oosting M, Smeekens SP, Jaeger M, et al. Integration of multi-omics data and deep phenotyping enables prediction of cytokine responses. Nat Immunol. 2018;19:776–86.
37. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-omics factor analysis–a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018;14:e8124.
38. López de Maturana E, Alonso L, Alarcón P, Martín-Antoniano IA, Pineda S, Piorno L, et al. Challenges in the integration of omics and non-omics data. Genes. 2019;10:238.
39. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, editors. Adv Neural Inf Process Syst 26 [Internet]. Curran Associates, Inc.; 2013 [cited 2019 Apr 30]. p. 3111–9. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.
40. Macaulay IC, Ponting CP, Voet T. Single-cell multiomics: multiple measurements from single cells. Trends Genet. 2017;33:155–68.
41. Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet. 2013;14:618–30.
42. Levitin HM, Yuan J, Sims PA. Single-cell transcriptomic analysis of tumor heterogeneity. Trends Cancer. 2018;4:264–8.
43. Winterhoff B, Talukdar S, Chang Z, Wang J, Starr TK. Single-cell sequencing in ovarian cancer: a new frontier in precision medicine. Curr Opin Obstet Gynecol. 2019;31:49–55.
44. Shalek AK, Benson M. Single-cell analyses to tailor treatments. Sci Transl Med [Internet]. 2017 [cited 2019 Apr 30];9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5645080/.
45. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet. 2015;16:133–45.
46. Kim K-T, Lee HW, Lee H-O, Song HJ, Jeong DE, Shin S, et al. Application of single-cell RNA sequencing in optimizing a combinatorial therapeutic strategy in metastatic renal cell carcinoma. Genome Biol. 2016;17:80.

Publisher's Note

Genomic Mutation Identification in Mice Using Illumina Sequencing and Linux-Based Computational Methods

John A. Williams,[1,2,3] George Powell,[1,4] Ann-Marie Mallon,[1] and Michelle M. Simon[1,5]

[1]MRC Harwell Institute, Mammalian Genetics Unit, Harwell Campus, Oxfordshire, United Kingdom
[2]Institute of Translational Medicine, University Hospitals Birmingham NHS Foundation Trust, Birmingham, United Kingdom
[3]Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, United Kingdom
[4]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom
[5]Corresponding author: *m.Simon@har.mrc.ac.uk*

Genetically modified mice are an essential tool for modeling disease-causing mechanisms and discovering gene function. SNP genotyping was traditionally used to associate candidate regions with traits in the mouse, but failed to reveal novel variants without further targeted sequencing. Using a robust set of computational protocols, we present a platform to enable scientists to detect variants arising from whole-genome and exome sequencing experiments. This article guides researchers on aligning reads to the mouse genome, quality-assurance strategies, mutation discovery, comparing mutations to previously discovered mouse SNPs, and the annotation of novel variants, in order to predict mutation consequences on the protein level. Challenges unique to the mouse are discussed, and two protocols use self-contained containers to maintain version control and allow users to adapt our approach to new techniques by upgrading container versions. Our protocols are suited for servers or office workstations and are usable by non-bioinformatics specialists. © 2019 by John Wiley & Sons, Inc.

Keywords: mouse genomics • mouse mutagenesis screens • mutation detection • single-nucleotide detection • variant annotation • whole genome sequencing

How to cite this article:
Williams, J.A., Powell, G., Mallon, A.-M., & Simon, M.M. (2019). Genomic mutation identification in mice using illumina sequencing and linux-based computational methods. *Current Protocols in Mouse Biology*, *9,* e64. doi: 10.1002/cpmo.64

INTRODUCTION

The first-ever completed human genome originally took around 12 years to sequence, and was completed in 2002; sequencing of the mouse genome was completed 1 year later. Both species were initially sequenced using Sanger sequencing, which was developed in the 1970s. Even though the technology was fast and efficient for its time, it was wholly unsuitable and not cost effective for whole-genome sequencing of multiple individuals. Nowadays, with the advent of next-generation sequencing (NGS), it is possible to complete whole-genome sequencing of an individual within 1 day in a very cost-effective and accurate manner. This revolutionary innovation has allowed the efficient detection

CURRENT PROTOCOLS
A Wiley Brand
Current Protocols in Mouse Biology e64, Volume 9
Published in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/cpmo.64

of mutations in humans, mice, and a plethora of other species to aid the development of personal medicines, which are emerging as a reality in today's world.

One of the primary reasons the mouse is used as a model organism of human disease is because the mouse genome shares a greater proportion of similarity with the human genome than one might expect given the species' outward appearance; approximately 99% of functional genes in the human genome have an orthologous counterpart gene in mice (Mouse Genome Sequencing Consortium, 2002). The sequence of such a gene is likely to be evolutionarily conserved (to an extent, particularly in protein-coding regions) between humans and mice. Thus, it is possible to use the mouse as a model to infer the functional effects that specific genes might have on the human phenotype. Of interest is: how are disease phenotypes manifested, and what genes affect them?

Two major approaches are adopted in the effort toward understanding the genetic components of disease, and these are the genotype-driven (reverse genetics) and phenotype-driven (forward genetics) methodologies. With the former, the idea is to target a specific gene of interest and assess its effect on the resultant phenotype of an organism, in this case the mouse. With the latter, it is the phenotype of the organism that is initially observed, and the aim is to deduce the causative gene. Both approaches require genetic mutation/manipulation, either spontaneous in nature or by intervention.

N-Ethyl-*N*-nitrosourea (ENU) is a common chemical mutagen used to confer single point mutations in mouse spermatogonia. These mutations are supposedly distributed at random across the genome and occur at an approximately 1,000-fold higher rate than spontaneous mutations (Concepcion, Seburn, Wen, Frankel, & Hamilton, 2004). The fact that ENU preferentially causes point (single-nucleotide) mutations means that the effects of small variations—such as minimally altered coding sequences—can often be discovered. Such variations might confer altered or reduced function of a gene's product. This enables a broad range of alleles (multiple allelic series) and phenotypes to be

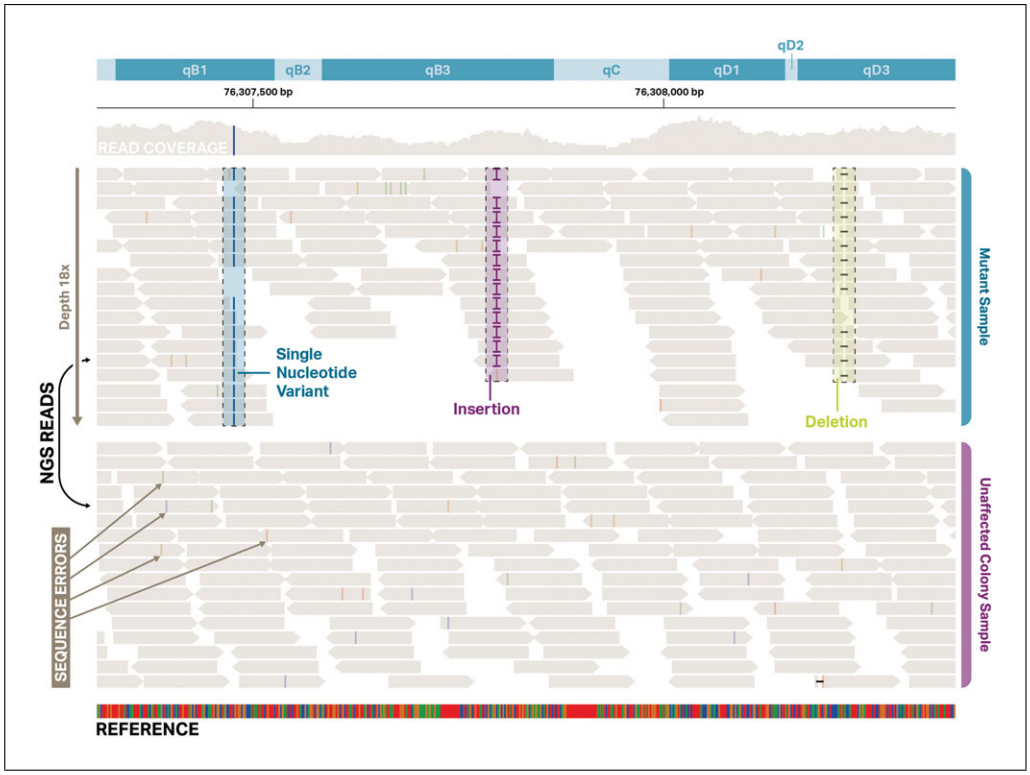


Figure 1 Graphical representation of NGS data including SNVs and small indels in a wild-type and mutant mouse sample.

observed and characterized (Qian, Mahaffey, Alcorn, & Anderson, 2011). These alleles include hypermorphic, amorphic, antimorphic, and neomorphic mutations (O'Brien & Frankel, 2004), resulting in a wide spectrum of manipulations and phenotypes.

Gene-driven screens are used to assess the phenotypes of mice after a known genetic lesion is produced. These screens can generate different mutation types using an array of different technologies. For example, the International Mouse Phenotyping Consortium (IMPC) originally used LacZ reporter methods to produce gene knock-outs in mice, while Panda et al. (2013) used transcription activator–like effector nucleases (TALENs) to produce mutations in mice. Increasingly more laboratories and large phenotype screens are using clustered regularly interspaced short palindromic repeats and the Cas-9 protein (CRISPR/Cas9) to generate direct mutations in the zygote. Due to the prior knowledge of the mutation type and genomic location, the use of NGS to map and identify novel causative mutations is usually not required. However, NGS still facilitates modern gene-driven screens, typically by quantifying the accuracy of the method (Mianné et al., 2016). For instance, in addition to the use of NGS to identify ad hoc mutations that may modify the phenotypes observed, NGS is typically used to verify any off-target genomic modifications widely seen in experiments that use CRISPR to induce mutations in mice. This protocol focuses on single nucleotide variations (SNVs) primarily but not exclusively produced by ENU mutagenesis screens, as shown in Figure 1.

STRATEGIC PLANNING OF MUTATION-DETECTION EXPERIMENTS

Forward genetic screens such as ENU mutagenesis can be either dominant or recessive. The dominant screen refers to mutations that show a dominant or semidominant phenotype in the first generation from the mutagenized male. The advantage here is that only one round of breeding is required before phenodeviants are identified; however, a disadvantage is that the causative mutations are identified less frequently, reflecting the high frequency of mutations in the first generation. To characterize recessive mutations, two more generations of breeding are required before a phenodeviant can be detected. Here, the G1 progeny are mated to a wild-type mouse, making the mutations in the G2 progeny heterozygous. To make these ENU-induced mutations homozygous and easier to characterize phenotypically, the G2 mice are intercrossed to produce G3 mice. Alternatively, female G2 mice are backcrossed to G1 males, again producing homozygous mutations (Balling, 2001). G1 dominant and G3 recessive breeding schemes are outlined in Figure 2. Typically, investigators will sequence the G3 mice to get the candidate SNVs for a particular phenotype. However, to obtain all the SNVs for the array of phenotypes generated from one mating, the G1 can be sequenced; then, all the SNVs will be captured and the G3 genotyped for particular SNVs or SNVs within a candidate region (Potter et al., 2016). Traditionally, candidate regions for a particular phenotype were obtained via positional cloning and fine mapping strategies (Justice, Noveroske, Weber, Zheng, & Bradley, 1999) to determine regions of homozygosity. With NGS and the protocols here, only gross mapping or indeed no mapping is required because sufficient low-confidence SNVs are filtered from further consideration to generate a concise list of plausible SNVs for validation. For this reason, it is advisable to have sufficient variant information from your mouse colony to use for comparison and filtering purposes. While a discussion of experimental validation techniques is beyond the scope of this article, validating genetic variants through Sanger sequencing assays remains a valuable tool to confirm putative mutations characterized using techniques covered here.

There are many common experimental platforms for performing NGS, each with their own chemistry, such as Illumina (Bennett, 2004), PacBio (Korlach et al., 2017), and others. A data variation that affects analysis is sequence read length. This article assumes that sequencing was performed with a version of the Illumina HiSeq platform to produce

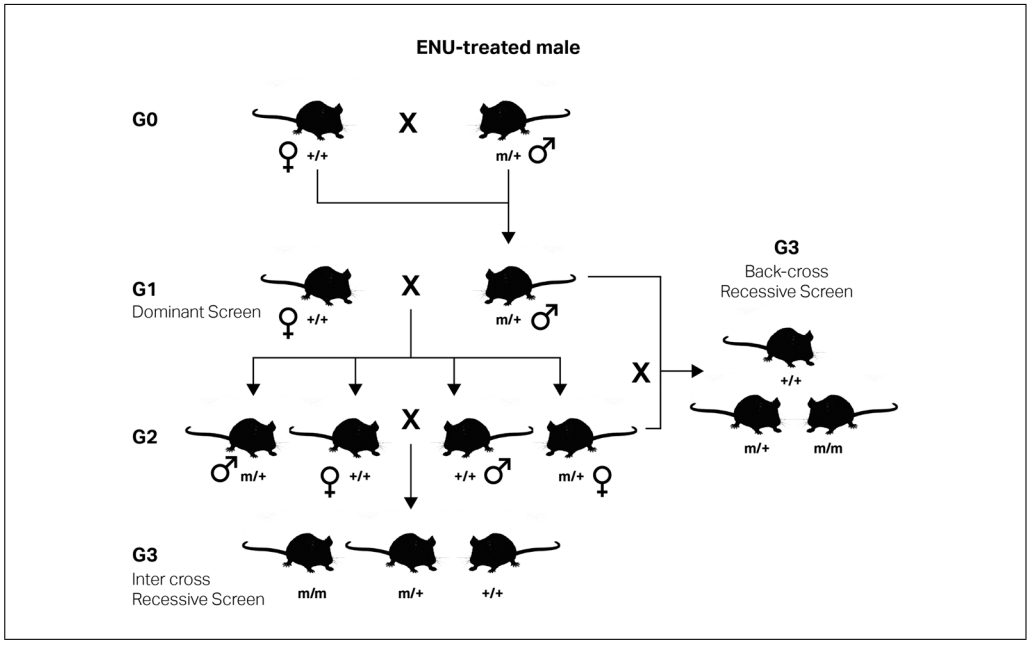


Figure 2 ENU-mutagenized males are mated to wild-type females. Each G1 offspring carries a unique set of mutations and therefore will exhibit different phenotypes. To segregate the phenotypes with a mutation or mutations, the G1 male mice are mated to wild-type females. The G2 offspring are then either intercrossed to each other or mated back to the original G1. The G3 progeny can be phenotyped for both recessive and dominant mutations. Typically, wild-type females of a different genetic background are used in the breeding scheme to facilitate the genetic mapping of any phenodeviant G3 offspring.

short reads of 75- to 100-base-pair (bp) length. Though sequencing quality can decrease with longer reads, especially toward the 3' end, the increase in read length improves mapping. Proper quality control (QC) can access the accuracy of longer (100-bp) reads. If needed, read trimming can be used to improve mapping. We recommend trimming reads after accessing quality of the FASTQC report generated in Basic Protocol 1, and use of TrimGalore for that purpose (Krueger, 2019; Martin, 2011). For a qualitative review of adapter trimming methods, see Fabbro, Scalabrin, Morgante, and Giorgi (2013). If other sequencing protocols are followed, the choice of aligner and pre-processing QC steps will need to be adjusted. If other sequencing protocols are followed, starting from Basic Protocol 1, step 17, the rest of the Basic Protocol 1 and subsequent protocols can be followed without adjustment provided that a SAM file of aligned reads can be supplied. NGS can produce either single-end or paired-end reads (Sengupta et al., 2011) from DNA fragments produced from a library of templates from the mouse genome. This article assumes that readers will be using paired-end reads, though Basic Protocol 1 may be modified to use single-end reads and the subsequent protocols are not dependent upon single- or paired-end design. While single-end reads sequence each DNA fragment once, paired-end reads sequence each twice, once in each direction starting at each end of the fragment. This produces twice the number of reads for the same amount of sequenced fragments. Paired-end sequencing is more costly than single-end; however, it offers several advantages by indicating positions of reads relative to each other. This facilitates more accurate identification of insertions and deletions (indels) than single-end reads, and increases reliability of SNPs called, especially in repetitive regions. For a reliable assessment of sequencing technology, see Park and Kim (2016); also see Shendure, Porreca, and Church (2008) for relevant reviews.

COMPUTATIONAL RESOURCE REQUIREMENTS

NGS data sets are large; thus, a large amount of computational resources are required to efficiently handle such data. A single sequencing experiment can produce several gigabytes of compressed reads, and extensive memory (RAM) is needed to process such volumes of data. Modern computers with at least 8 GB of RAM and 500 GB of disk space are sufficient for small analyses, provided that a modern 64-bit processor and multiple cores are included. This tutorial assumes a Unix-like environment, either Linux or a recent version of Mac OS. Windows users may be served by using a virtual machine to emulate a Linux distribution, as many software tools used in this article are not supported in Windows environments. Having multiple physical or virtual cores will allow parallel processing at certain steps, increasing processing speed dramatically. While such a configuration on a personal workstation may suffice, in practice, servers or cloud computing environments are utilized. An introduction to using remote servers is beyond the scope of this article, which has been written to be easily implemented either locally on a workstation, or remotely on a server directly, without need to use a grid engine to submit work to distributed computing resources.

ALIGNMENT OF SEQUENCING READS AND QUALITY CONTROL

BASIC PROTOCOL 1

This protocol describes how to align sequences to the current mouse reference genome. Starting with some housekeeping setup to make this and following protocols accessible in any Unix-like computing environment, the current reference genome is acquired and processed with SAMtools (Li et al., 2009). The FASTQC tool is used to access the quality of the NGS reads themselves before initiating alignment (Andews, 2015). Sequencing reads are then aligned to the genome using the Burrows-Wheeler Aligner (BWA; Li & Durbin, 2009). Picard Tools is used to mark duplicate reads in the file, and SAMtools is again used for additional postprocessing (Picard Toolkit, 2018). The created files are used in Basic Protocol 2 to detect variants.

Materials

Hardware

Workstation (Mac OS X, Linux, or any Unix-like system) with 500 GB of RAM, a 64-bit processor with at least two real or virtual cores, and 500 GB free disk space either on the workstation or in an external drive. An alternative is a remote server meeting at least these specifications. Administrator privileges are required for initial software setup, but do not have to belong to the end user. See Computational Resource Requirements for more information.

Software

A Java Runtime Environment (JRE) can be verified by visiting *https://www.java.com/en/download/installed.jsp?detect=jre*.
A Docker platform installation (Merkel, 2014)
FASTQC version 0.11.18 (Andews, 2015)
SAMtools version 1.9 (Li et al., 2009)
GATK version 4.0.11.0 (McKenna et al., 2010)
BWA version 0.7.17 (Li & Durbin, 2009)

Unix tools

The following bash commands are used, which should be installed on any modern Linux or Mac OS X distribution:
wget, tar, sed, cat

Datasets

Raw sequencing reads in FASTQ format from a sequencer. In absence of user-generated data, the Sequence Read Archive (*https://www.ncbi.nlm.nih.gov/sra*) may be investigated for test data (Leinonen, Sugawara, Shumway, & International Nucleotide Sequence Database Collaboration, 2011).

Setup software and file paths

In this and the following protocols, a working directory must be used. Software can likewise be stored in any accessible directory. Setting up one-time aliases for these directories enables seamless integration with this protocol. See below to set up aliases. If the software directory is `/NGS/Software/MouseMutations`, create a variable "`mySoftware`" to store this location. If the working directory is `/NGS/working_projects/MouseMutations`, create a variable "`myProjects`" to store this location.

1. Set up aliases for your directories by typing:

```
mySoftware=/NGS/Software/MouseMutations
myProject=/NGS/working_projects/MouseMutations
```

   Once the data are obtained in FASTQ format, the initial step is to align the sequence reads to the genome. Before starting this endeavor, it is useful to access the quality of the sequences themselves. FASTQC is a community-adopted program to access sequence quality. FASTQ files are stored in the "`myProject`" directory and labeled as:

```
sample_1.fastq
sample_2.fastq
```

Perform quality control

2. Each FASTQ file is from one of the "paired ends" of a single experiment. If experimental protocols use single-end reads, each sequencing run will produce one `sample.fastq` file instead of two. The `_1` and `_2` labels denote files, which are paired ends of the same experiment. This protocol may be modified to use single-end reads by typing "`sample.fastq`" instead of both "`sample_1.fastq`" and "`sample_2.fastq`" in steps 2 and 16 of this protocol. Run the following command to execute FASTQC on both of the FASTQ files. This will output reports in FASTQ format (html files and zip files). Adding "`--threads 8`" allows the program to use eight computational threads to speed up processing. Change this to match the settings on the workstation in use. To access how many threads are available on a Linux environment, type:

```
nproc
```

   Or on a Mac OS X environment type:

```
sysctl hw.logicalcpu
```

Output will inform the code below, as well as any further decisions on the number of threads or processes available for use. FASTQC reports will be generated and may be interpreted by viewing documentation in "Internet Resources."

```
$mySoftware/FastQC/fastqc *fastq --threads 8
```

Obtain and index the mouse genome

To conveniently obtain the mm10 genome from mouse strain C57BL/6J, Illumina creates builds by Ensembl, UCSC, and the NCBI. The Mouse Genomes Project contains draft assembled genomes for 16 laboratory- and wild-derived strains, which may be used in place of the standard mm10 genome as biological questions dictate. These may be accessed at *https://www.sanger.ac.uk/science/data/mouse-genomes-project* (Adams, Doran, Lilue, & Keane, 2015).

3. Download the mm10 UCSC genome:

```
wget ftp://igenome:G3nom3s4u@ussd-ftp.illumina.com/
Mus_musculus/UCSC/mm10/Mus_musculus_UCSC_mm10.tar.gz
```

4. Uncompress the files:

```
tar -zxvf Mus_musculus_UCSC_mm10.tar.gz
```

5. Change to the `Chromosomes` directory:

```
cd Mus_musculus/UCSC/mm10/Sequence/Chromosomes/
```

6. Edit a header in the `.fa` files downloaded to remove "`chr`":

```
for f in *.fa; do sed -i 's/chr//g' $f; done;
```

7. Combine individual chromosomes with the "`cat`" command into one output "`ref_mm10.fasta" file`:

```
cat chr1.fa chr2.fa chr3.fa chr4.fa chr5.fa chr6.fa
chr7.fa chr8.fa chr9.fa chr10.fa chr11.fa chr12.fa
chr13.fa chr14.fa chr15.fa chr16.fa chr17.fa chr18.fa
chr19.fa chrM.fa chrX.fa chrY.fa > ref_mm10.fasta
```

8. Move the `.fasta` file into a new "`musRefs`" directory and move back to the project directory:

```
mkdir -m 777 $myProject/musRefs_10/
mv ref_mm10.fasta $myProject/musRefs_10/
cd $myProject
```

9. Index the mm10 genome for use with BWA:

```
$mySoftware/bwa/bwa index -a bwtsw musRefs_10/
ref_mm10.fasta
```

10. Enable permissions to edit the files in musRefs_10:

```
cd musRefs_10; chmod 777 *
```

11. Index the mm10 genome for use with SAMtools:

```
$mySoftware/samtools/bin/samtools faidx ref_mm10.fasta
```

12. Index the mm10 genome for use with the Genome Analysis Toolkit (GATK). If not previously downloaded, pull the following GATK image by typing:

```
docker pull broadinstitute/gatk:4.0.11.0
```

13. Then, mount the GATK image and re-create the `myProject` variable:

```
docker run -t -i -v $PWD:/gatk/MouseMutations
broadinstitute/gatk:4.0.11.0 /bin/bash
myProject=/gatk/MouseMutations
```

14. Use GATK to create the index:

```
$mySoftware/gatk-4.0.11.0/gatk CreateSequence
Dictionary --REFERENCE=$myProject/ref_mm10.fasta
--OUTPUT=$myProject/ref_mm10.dict
```

15. Move back out of the Docker container and into the `$myProject` directory:

```
exit; cd $myProject
```

Align experimental reads to the mouse genome and format results

16. With the mm10 reference genome FASTA file and the two mouse FASTQ files given below, run BWA to align the genome to mm10. In the command below, the BWA program is installed in the `$mySoftware/bwa/bwa` directory. "`mem`" is called to align 70 bp to 1 Mbp-length reads with "`-t`" of 30 threads. Adjust the number of threads to fit the workstation. The FASTQ file has been indexed as above and is stored in `musRefs_10/ref_m10.fasta`. As stated above, paired-end sequencing has produced two FASTQ files, one for each pair. If single-end sequencing has been used, simply give one FASTQ file. BWA assumes that the two FASTQ files given are mates, with one FASTQ file being interpreted as "`reads.fastq`" and the other as the corresponding "`mates.fastq`" pair. See the BWA documentation for

more information. Run the line below. The end result will be an aligned "SAM" file in the `$myProject` directory.

```
$mySoftware/bwa/bwa mem -t 30 musRefs_10/ref_mm10.fasta
sample_1.fastq sample_2.fastq > sample.sam
```

17. Use GATK's included version of Picard Tools (Picard Toolkit, 2018) to add read group names to the SAM file created above. Then mount the Docker image to use GATK, including a map to the local `$myProject` directory which you are currently in. Once there, re-create the `$myProject` variable in the Docker container:

```
docker run -t -i -v $PWD:/gatk/MouseMutations
broadinstitute/gatk:4.0.11.0 /bin/bash
myProject=/gatk/MouseMutations
```

18. Now, use Picard Tools to allow GATK's HaplotypeCaller to work correctly. The values passed to number a read group (`--RGID`), a read library (`--RGLB`), and platform (`--RGPL`), a platform unit or run barcode (`--RGPU`), and a sample name (`--RGSM`) can be either arbitrary or have assigned meaning.

```
gatk AddOrReplaceReadGroups --INPUT=$myProject/
sample.sam --OUTPUT==$myProject/sample.named.sam
--RGID=1 --RGLB=lib1 --RGPL=illumina --RGPU=unit1
--RGSM=1
```

Sort the aligned reads

19. Next, sort the SAM file by leftmost coordinates to allow Picard Tools to correctly find duplicates, and compress the SAM file into its binary analog, a BAM file. This is done using Picard Tools bundled with GATK and executing the `SortSam` command. The output of this command is a sorted BAM file. Do this in one step as shown below.

```
gatk SortSam --INPUT==$myProject/sample.named.sam --
OUTPUT==$myProject/sample.sorted.bam --SORT_ORDER=
coordinate
```

Deduplicate aligned, sorted reads

20. To mark potential PCR artifacts such as duplicate reads, run the MarkDuplicates Picard Tools function in GATK. The input "`INPUT`" flag is assigned to the sorted BAM above. The output of the new file "`OUTPUT`" ends in `sorted.rmDub.bam`. The MarkDuplicates command requires a metrics file to be included to output statistics; call this `Input.tmp_file`. Ask MarkDuplicates to remove duplicates found via the `REMOVE_DUPLICATES=true` option, and to assume the file is sorted (`ASSUME_SORTED`). Lastly, set the VALIDATION_STRINGENCY to silent to allow processing of the entire file without the program failing if errors are encountered. Use a "`tmp/`" directory as your temporary "TMP_DIR" to store files, which will be deleted after the completion of the protocol.

```
gatk MarkDuplicates --INPUT==$myProject/sample.
sorted.bam --OUTPUT==$myProject/sample.sorted.
rmDup.bam --METRICS_FILE=Input.tmp_file
--REMOVE_DUPLICATES=true --ASSUME_SORTED=true
--VALIDATION_STRINGENCY=SILENT --TMP_DIR=tmp/
```

21. Exit the Docker container by typing:

```
exit
```

    to return to the `$myProject` directory in the native shell (set above as `/NGS/working_projects/MouseMutations`).

Index the deduplicated reads

22. Lastly, index the deduplicated BAM file with SAMtools to run GATK options in Basic Protocol 2. This will create an index file in the `$myProject` directory.

```
$mySoftware/samtools/bin/samtools index
$myProject/sample.sorted.rmDup.bam
```

    Now the sorted, deduplicated, indexed BAM file is ready to be analyzed with the GATK.

BASIC PROTOCOL 2

MUTATION DETECTION AND VARIANT GENOTYPING

This protocol focuses on using the GATK for variant discovery and genotyping. While different tools can be used for this protocol, GATK is in current development and well documented. Of note, previous versions of GATK included custom-built multiprocessing ability. New versions (>4.0.8) use Spark (Zaharia, Chowdhury, Franklin, Shenker, & Stoica, 2010) to enable multiprocessing. However, this is under development and should not be used for production work until moved out of "Beta" release, as results via Spark currently do not match those shown below in this protocol. Input for this protocol is the SAM file from Basic Protocol 1. First, the Haplotype Caller is used to build a de novo assembly of haplotypes in regions of higher-than-expected variation. This facilitates calling indels and SNPs that may be close together or in high-complexity regions. Next, the re-assembled haplotypes are used as input for calling genotypes with the Genotype Caller to produce a Variant Call Format (VCF) file. Lastly, this VCF file is annotated based on specific QC hard filters and is then ready to be compared to known mutations in Basic Protocol 3.

Materials

Hardware

Workstation (Mac OS X, Linux, or any Unix-like system) with 500 GB of RAM, a 64-bit processor with at least two real or virtual cores, and 500 GB free disk space either on the workstation or in an external drive. An alternative is a remote server meeting at least these specifications. Administrator privileges are required for initial software setup, but do not have to belong to the end user. See Computational Resource Requirements for more information.

Software

A JRE can be verified by visiting *https://www.java.com/en/download/installed.jsp?detect=jre*
GATK version 4.0.11.0
Docker

Datasets

The sorted, deduplicated, indexed BAM file from Basic Protocol 1:

```
$myProject/sample.sorted.rmDup.bam
```

The indexed mouse reference genome from Basic Protocol 1 along with its index and dictionary:

```
$myProject/musRefs_10/ref_mm10.fasta
$myProject/musRefs_10/ref_mm10.dict
$myProject/musRefs_10/ref_mm10.fasta.fai
```

1. First, use the Haplotype Caller in GATK to call local haplotype assemblies for your sorted, deduplicated BAM file. First, load the Docker image and mount the `$myProject` directory as in Basic Protocol 1:

```
docker run -t -i -v $PWD:/gatk/MouseMutations
broadinstitute/gatk:4.0.11.0 /bin/bash
myProject=/gatk/MouseMutations
```

2. Use the built in GATK script to call the `GenomeAnalysisTK.jar` file (the GATK program) within the GATK Docker image. `--java-options "-Xmx4g"` indicate that you are giving 4 GB of memory to the JRE. GATK is using the HaplotypeCaller program, and the "`-R`" flag indicates the directory of your reference FASTA file for the mouse genome. The "`-I`" flag points to your sorted deduplicated BAM file. The `-O` flag indicates where the output will be stored (a VCF file). Lastly, the "`-ERC`" GVCF flag indicates that the algorithm will output an intermediate genomic VCF. The resulting output file will be `$myProject/sample.sorted.rmDup.g.vcf`, a genomic VCF fie. While superficially the same as a VCF file that will be will encountered later, this also encodes information from the reference genome (`-R`) and indicates both variant site records and blocks of nonvariant sites. Type:

```
gatk --java-options "-Xmx4G" HaplotypeCaller -R
$myProject/musRefs_10/ref_mm10.fasta -I
$myProject/sample.sorted.rmDup.bam -O
$myProject/sample.sorted.rmDup.g.vcf -ERC GVCF
```

Once this file is created by the HaplotypeCaller, it can be genotyped with the GenotypeGVCFs function. Simply pass the new "`g.vcf`" file to this function as below. Here, call GATK as above, but specify the GenotypeGVCFs function. The reference FASTA is the same as above, and the input is your "`g.vcf`" from the Haplotype Caller. The output is "`toFilter-gatk.vcf`," which will be used for variant filtering.

3. Type:

```
gatk GenotypeGVCFs -R $myProject/musRefs_10/ref_
mm10.fasta -V $myProject/sample.sorted.rmDup.g.
vcf -O $myProject/toFilter-gatk.vcf
```

4. Various filters can be applied. It is recommended to filter by read depth and sequencing quality. While in humans GATK's VQSR (variant recalibration based on 1000 Genomes and other training data) can be used, in mouse no such models exist due to lack of training data. Thus, hard filters may be applied. The expected result of this step is to encode the seventh field, FILTER, of `toFilter-gatk.vcf` with either the string "`PASS`" or another code indicating that the SNP should be filtered out early in Basic Protocol 3. The filter below will filter based on a quality score of 30 and a read depth for each SNP of at least 10×. For variants that do not pass this simple filtration, instead of "`PASS`," the "`--filter-name`" of "`myDepthQualityFilter`" will be input. Output will be a VCF file with these filter options added.

```
gatk VariantFiltration -R $myProject/musRefs_10/
ref_mm10.fasta -V $myProject/toFilter-gatk.vcf -
O $myProject/filtered-gatk.vcf --filter-expression
"QUAL > 30.0 && DP == 10" --filter-name "myDepth
QualityFilter"
```

5. Finally, exit the Docker container to return to the rest of the protocol. Continue with Basic Protocol 3 unless Alternate Protocol or Support Protocol 2 apply. Alternate Protocol addresses experiments with multiple biological replicates, a common situation. While this protocol can be repeated for each replicate, Alternate Protocol automates the process. Support Protocol 2 provides for additional "hard filtering" steps, which may be needed if the quality of the experimental data is suspect, and provides a template for adding additional filtering steps.

```
exit
```

BASIC PROTOCOL 3

NOVEL MUTATION DISCOVERY

After detecting mutations with the GATK pipeline described above, we wish to compare our findings to known external results, and therefore annotate only novel findings. If annotating all mutations is desirable (and one does not want to annotate only novel mutations), this protocol may be skipped, and the VCF file from Basic Protocol 2 can be used as input to Basic Protocol 4. This protocol will produce a file with unknown variants. We will then annotate unknown variants in Basic Protocol 4.

Materials

Hardware

Workstation (Mac OS X, Linux, or any Unix-like system) with 500 GB of RAM, a 64-bit processor with at least two real or virtual cores, and 500 GB free disk space either on the workstation or in an external drive. An alternative is a remote server meeting at least these specifications. Administrator privileges are required for initial software setup, but do not have to belong to the end user. See Computational Resource Requirements for more information.

Unix tools

The following bash commands are used, which should be installed on any modern Linux or Mac OS X distribution:

wget, gunzip, awk, cut, comm

Datasets

The filtered GATK processed VCF file from Basic Protocol 2 (or the alternate or support protocols, as mentioned above):

```
$myProject/filtered-gatk.vcf
```

A reference panel of known SNPs downloaded from the Sanger Institute (see below) will be acquired during this protocol. Sanger provides current data updated via the URL *ftp://ftp-mouse.sanger.ac.uk/current_snps*.

1. Make sure to be in the myProject working directory:

```
cd $myProject
```

2. Download the mouse known mutation file with wget:

```
wget ftp://ftp-mouse.sanger.ac.uk/current_snps/
mgp.v5.merged.snps_all.dbSNP142.vcf.gz
```

3. Unzip the mouse mutation file. This will output one file, `mgp.v5.merged.smps_all.dbSNP142.vcf`:

```
gunzip mgp.v5.merged.snps_all.dbSNP142.vcf.gz
```

Note that in practice every mouse colony is different and may contain its own specific mutations due to drift. Laboratories combine the file above with known mutations from their own colony to filter out variants they have previously characterized. It is suggested that current researchers do the same. Going forward, assume that the unzipped file above contains combined mutations from one's own lab and the known SNPs database.

4. Next, remove the words "`chr`" from the "`$myProject/filtered-gatk.vcf`" file in column 1 of the VCF file, and remove all headers, using the "grep" tool. Then, filter each field for only variants that passed filters from the GATK protocol using "awk." To clean up output for comparing this merged dpSNP file to the filtered GATK file, "cut" the fields 1, 2, 4, and 5, then "sort" them into a temporary file called "a." Field 3, which holds SNP rs IDs, will be withheld and re-annotated in Basic Protocol 4:

```
grep -v '^#' mgp.v5.merged.snps_all.dbSNP142.vcf |
awk '$7 == "PASS"' | cut -f1,2,4,5 | sort -u > a
```

5. Next, repeat step four (above) with the `gatk.vcf` file, creating a temporary file called "b":

```
grep -v '^#' filtered-gatk.vcf | awk '$7 == "PASS"' |
cut -f1,2,4,5 | sort -u > b
```

Table 1 The Structure of Mutant.vcf Passed to VEP for Annotation[a]

Field position	Field name	Field meaning	Example
1	CHROM	Chromosome number	2
2	POS	Genomic coordinate	112876894
3	ID	SNP ID	rs246375767
4	REF	Reference allele	C
5	ALT	Alternate allele	T

[a]Output from the "`mutant.vcf`" file explained. Each field can be used by VEP and other resources to characterize known and novel variants. The ID field is marked as "." if the SNP ID is unknown. Indels, CNVs, and other mutation types are characterized by the same file format and can also be input in to variant prediction protocols (see Basic Protocol 4).

6. Then, compare the files with the "comm" utility. The comm flag "`-13`" excludes lines unique to file "`a`" with flag 1 (the dpSNP reference) and in common to both (flag 3), keeping only field 2 of standard comm output, the lines unique to only "`b`," and the formatted `filtered-gatk.vcf` file holding unique mutations called `novel_snps`:

```
comm -13 a b > novel_snps
```

7. Lastly, `novel_snps` must be formatted to restore the lost column "3" cut out in steps 4 and 5, using a "`.`" to indicate yet-unknown SNPs. These will be annotated with Ensembl-sourced rsIDs in Basic Protocol 4. This line produces `mutant_vcf`, which is the input for Basic Protocol 4:

```
awk -F' ' '{$2 = $2 FS ".”; print $1," ",$2," ",$3,
" ",$4}' novel_snps > mutant.vcf
```

The resulting `mutant.vcf` file contains potentially novel SNVs. The five fields referenced above are explained in Table 1.

The only input needed for Basic Protocol 4 is the resulting file, `mutant.vcf`. This protocol involves downloading reference databases specific to your strain of mouse automatically. Characterize by Basic Protocol 4.

BASIC PROTOCOL 4

VARIANT INTERPRETATION

To characterize genomic variants identified as unique to our experiment above, we employ the Variant Effect Predictor (VEP) from Ensembl (McLaren et al., 2016). VEP is available in many formats, but to maintain compatibility it is becoming best practice to use a Docker instance. This ensures that VEP and all its dependencies are managed in a self-contained ecosystem and do not interfere with any other programs you may have installed on your system.

Materials

Hardware

Workstation (Mac OS X, Linux, or any Unix-like system) with 500 GB of RAM, a 64-bit processor with at least two real or virtual cores, and 500 GB free disk space either on the workstation or in an external drive. An alternative is a remote server meeting at least these specifications. Administrator privileges are required

for initial software setup, but do not have to belong to the end user. See Computational Resource Requirements for more information.

Software

Docker version 18.06 or greater
The Ensembl Variant Predictor Docker container (version 95.1)

Datasets

The `mutant.vcf` file produced in Basic Protocol 3:

```
$myProject/mutant.vcf
```

Download VEP

1. Clone VEP Docker image and run VEP. Cloning the Docker image will take time initially and does not need to be done repeatedly. Run:

   ```
   docker pull ensemblorg/ensembl-vep:release_95.1
   ```

   Run the following to call up the help message (optional), which gives parameters useful for diagnosing problems and expanding annotation options:

   ```
   docker run -t -i ensemblorg/ensembl-
   vep:release_95.1./vep --help
   ```

2. After VEP is installed, create a directory (folder) in which to put the `mutant.vcf` file:

   ```
   mkdir vep_data
   ```

3. Then, change the permissions to make sure that you can read, write, and execute programs in the directory.

   ```
   chmod a+rw vep_data
   ```

4. Now, move into that directory and copy the "`mutant.vcf`" file there:

   ```
   cd vep_data
   mv $myProject/mutant.vcf.
   ```

5. To make the following steps easier and in line with the official VEP tutorial, create a temporary variable in the bash shell which will locate the `VEP_data` folder. Using `$PWD` (below) ensures that this code snippet will work wherever one is in the workstation directories.

   ```
   myAnnotation=$PWD
   ```

Install VEP application programming interface (API)

The Docker image contains everything needed to run VEP. To install VEP within the Docker container, run the perl `INSTALL.pl` script contained in the Docker image. To do

this, simply enter the command under step 6. The "`t`" flag indicates that a pesudo-TTY is enabled so Docker can pass the image signals, the "`i`" enables the user to add information to the Docker image via standard input, and the "`v`" flag binds the "`$myAnnotation`" directory, where the "`mutant.vcf`" file is stored, to the "`/opt/vep/.vep`" hidden directory in the VEP Docker image. Doing this allows local files to be stored and accessed via Docker, and allows files generated in the ongoing analysis to be locally accessible after closing the VEP Docker image down.

6. Then type:

```
docker run -t -i -v $myAnnotation:/opt/vep/.vep
ensemblorg/ensembl-vep:release_95.1 perl INSTALL.pl
```

7. Next, a prompt will appear asking to install the API. If this is the first time using VEP, type "`y`." If not, type "`n`" unless upgrading to version 95.1 of the API:

```
Do you want to continue installing the API (y/n)? y
```

Install the mouse genome annotation and FASTA files

8. If installing the API (or upgrading to a new version), type "`y`" to proceed with overwriting the current cache directory. You will also be asked if you wish to cache any files. To cache the mouse genome variant information, type "`y`" at the prompt as depicted below:

```
Do you want to install any cache files (y/n)? y
```

   The mouse VEP annotation is listed in version 95.1 of VEP as 145:

```
145: mus_musculus_vep_95_GRCm38.tar.gz
```

9. Type `145` at the prompt, and it will install the mouse VEP v94 annotation files.

10. Having the files stored locally on the workstation or server will ensure a quick analysis time. Likewise, installing the human annotations may be useful for any comparisons of mouse SNP consequences to human data. Downloading and processing VEP cache data will take some time, but dramatically speed up future analyses, and need only be done once. VEP will then ask if it should install FASTA files. This is not required, but is suggested if this protocol is to be used frequently. Type "`y`" at the prompt to install FASTA files:

```
Do you want to install any FASTA files (y/n)? y
```

    There are several mouse strains sequenced, with option 82 being the default mm10 reference genome on strain C57BL/6J, listed as mus_musculus. Other mouse genomes are available from *ftp://ftp.ensembl.org/pub/release-95/fasta/*.

```
82: mus_musculus
```

11. Type `82` at the prompt to download Mus_musculus FASTA files at the question mark "`?`".

```
? 82
```

Install third-party plugins

The VEP installer will then ask if any plugins should be installed. Again, these are optional, but type "`y`" to install, then browse the list as required. They are useful if using offline mode, which can help troubleshooting when acquiring data via secure connections.

```
y
```

12. Type `0` at the next prompt next to the question mark "`?`". This will install all available plugins for mouse data, which are useful for giving various computational predictions of variant effects. If some installations fail, do not panic; they are still optional.

```
? 0
```

Annotate variants

Finally, VEP is fully ready to use. The Docker image will exit, presenting the working directory "`myAnnotation`." You will recall that the `mutant.vcf` file was previously moved to the `$myAnnotation` directory.

13. Run the Docker image and load the bash prompt by typing:

```
docker run -t -i -v $myAnnotation:/opt/vep/.vep
ensemblorg/ensembl-vep:release_95.1 /bin/bash
```

14. Running this command brings up a new bash prompt. To locate the `$myAnnotation` directory which we mapped above to `/opt/vep/.vep`, run the following command:

```
vep@29f222ab2c50:~/src/ensembl-vep$ ls /opt/vep/.vep/
```

The mutant file, plugins, and the mouse reference cache will each be seen and be ready for analysis.

15. Next, to annotate the "`mutant.vcf`" file, run the code below. This runs VEP with the "`mutant.vcf`" as input, with the cache of data and running all options. See Table 2 for explanation of flags and output:

```
./vep -i /opt/vep/.vep/mutant.vcf --cache
--everything --species mus_musculus --output_file
/opt/vep/.vep/mutant_VEP.txt --stats_file /opt/
vep/.vep/mutant_VEP.html -offline
```

Table 2 Output Options Passed to VEP[a]

Option name	Use
`--cache`	Uses cache FASTA mouse data
`--everything`	Runs all available analyses and plugins
`--species`	Species mouse species
`--output_file`	Creates a text output file in `/opt/vep/.vep/mutant_VEP.txt`
`--stats_file`	Creates a web page with summary statistics in `/opt/vep/.vep/mutant_VEP.html`
`--offline`	Uses our cache and plugins without online access

[a]Output options for VEP with explanations. Note that after `--species`, one types a Latin or English name. If `--offline` is set, then the species name must be in Latin, as in `--species mus_musculus`.

An error message will indicate indicating "`INFO: disabling PolyPhen`." This is normal, as the polyphen option is only relevant for human. These options produce several files, which are explained in Table 2.

A section of output from VEP annotation is depicted below in text format (the standard output, in this case being "`mutant_VEP.txt`"). An additional output file is an interactive website file, "`mutant_VEP.html`." This file can be loaded in any web browser. Both of these files will be in the "`$myAnnotation`" directory. The `mutant_VEP.txt` file will contain several lines of headers, each of which explains a field in the tab delimited file. The first 13 fields describe the variant, its position in the genome, and any coding consequences to the allele and gene in which it inheres. The last column, number 14, describes many potential sources of data which are explained in the headers of the file.

As is evident on inspection of the headers in `mutant_VEP.txt`, there are many header flags (rows starting with ##), as are common with VCF output files. Scrolling through these headers reveals the purpose of comments in the "Extra" column, where annotations are separated by "`;`" semicolons. Otherwise, fields are tab separated.

After giving the `--everything` flag above, VEP annotates SNPs with many tools downloaded. VEP annotations with the `--everything` flag are described in Table 3.

It is worth noting that using the "`--everything`" flag is the most convenient way to retrieve annotations for mouse. However, the default VEP output is largely meant for human data, and flags such as "`-af`" called by "`--everything`" will include the allelic frequency in ethnic populations reported in the 1000 Genomes Project. These are not relevant to mouse work, but it is advised to keep them in the workflow in case there is a need to annotate human SNPs.

A cleaner way to access variants on a global scale is to look at the `mutant_VEP.html` file. It contains summary statistics on mutation consequences, coding consequences, variants by chromosome, and position in the coding protein. Figure 1 shows a screenshot of that output. Note that the output generated in this protocol is from a sample of the `mutant.vcf` file of parts of the "X" and "Y" chromosome. We see that most variants are either intergenic or intronic.

Table 3 VEP Annotates SNVs with Several Resources[a]

Option name	Use
`--sift b`	AA substitution via homology
`--polyphen b`	AA substitution effect on structure
`--ccds`	Adds CCDS transcript ID
`--uniprot`	Adds Uniprot ID
`--hgvs`	Adds HGVS ID
`--symbol`	Adds gene symbol
`--numbers`	Adds exon/intron numbers
`--domains`	Adds overlapping domains
`--regulatory`	Overlapping regulatory regions
`--canonical`	Canonical transcript
`--protein`	Ensembl protein ID
`--biotype`	Biotype of transcript
`--uniprot`	Uniprot ID
`--tsl`	Transcript support level
`--appris`	Isoform annotation
`--gene_phenotype`	Phenotype of overlapped gene
`--pubmed`	Pubmed IDs
`--variant_class`	Sequence ontology variant class

[a]All options that VEP will call if given the `--everything` flag. Note that PolyPhen is only available for humans, and SIFT can also be species-specific but will work for mouse.

16. Scroll through the downloaded web page, which can open in any modern web browser. While data based on the SNVs and variants in real mice will be different, the analysis format will be similar. To work with the text file itself, the best approach is to use the command line. Opening the output of VEP may not be feasible in a modern word processor, especially in a program such as MS Word, due to memory constraints. We recommend using the text editor "nano," installed on all Mac and Linux systems. To peruse the text file, one may search for a gene or SNP of interest using the "grep" command. To use grep to search for the gene Vipr2, type:

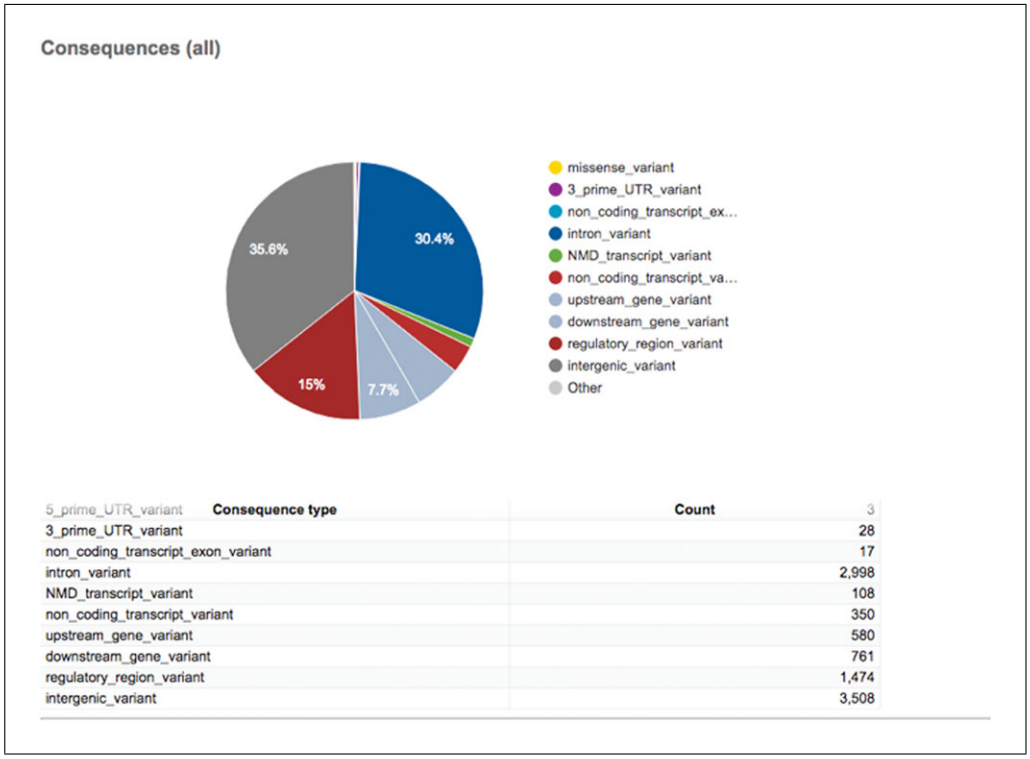
```
grep Vipr2 mutant_VEP.txt
```

If any variants are annotated to that gene, they will appear on the screen. VEP also includes a filter function. To explore this capability, type the following within the VEP Docker to display the help message:

```
./filter_vep --help
```

17. Lastly, exit the Docker image. Type:

```
exit
```



5_prime_UTR_variant	Consequence type	Count	3
3_prime_UTR_variant			28
non_coding_transcript_exon_variant			17
intron_variant			2,998
NMD_transcript_variant			108
non_coding_transcript_variant			350
upstream_gene_variant			580
downstream_gene_variant			761
regulatory_region_variant			1,474
intergenic_variant			3,508

Figure 3 A screenshot of the html web page generated by VEP displayed in a web browser. Note that tables displayed above are interactive, and you may scroll to find more information. The summaries generated are useful for overall characterization of mutations.

This ends Basic Protocol 4. Expected output relevant for further analysis are two files produced by VEP described above (Fig. 3). The html file provides a graphical overview of mutations and effects, while the annotated VCF file can be integrated into many analysis tools and has been analyzed for many consequences, as described in Table 3.

SUPPORT PROTOCOL 1

INSTALLATION OF NEEDED SOFTWARE

This protocol is written to aid in the installation of software needed for the analysis detailed in this article. Full instructions and help can be obtained from the websites referred to throughout the article and in the references provided.

Materials

Hardware

Workstation (Mac OS X, Linux, or any Unix-like system) with 500 GB of RAM, a 64-bit processor with at least two real or virtual cores, and 500 GB free disk space either on the workstation or an external drive. An alternative is a remote server meeting at least these specifications. Administrator privileges are required for initial software setup, but do not have to belong to the end user. See Computational Resource Requirements for more information.

Software

A JRE can be verified by visiting *https://www.java.com/en/download/installed.jsp?detect=jre*

Docker

Install Docker

1. To install Docker on a Linux system (assuming Ubuntu or Debian is used), type:

```
$ sudo apt-get update; apt-get install docker-ce;
```

   This will update the cache of available programs and install the community edition of Docker. Type the above command in the shell/terminal of choice, and a current version of Docker will be installed. This command first updates your cache and then installs docker-ce. If you have older versions of Docker installed, it is advisable to delete them first. Further guidance can be found at:

   https://docs.docker.com/install/linux/docker-ce/ubuntu/.

   On a system running Mac OS X, Docker can be installed by visiting *https://hub.docker.com/editions/community/docker-ce-desktop-mac* and following instructions.

Install FASTQC

2. To install FASTQC, first change to the "`$mySoftware`" directory by running:

```
cd $mySoftware
```

3. Then, use wget to download the compressed zip file:

```
wget https://www.bioinformatics.babraham.ac.uk/
projects/fastqc/fastqc_v0.11.8.zip
```

4. Lastly, unzip the downloaded zip file via the "unzip" utility pre-installed with Linux or Mac OS X:

```
unzip fastqc_v0.11.8.zip
```

Install BWA

5. To install the BWA software, pre-compiled binaries are available for many systems. An alternative approach is to use the "git" utility. Assuming a Linux Ubuntu environment, install git with:

```
sudo apt-get update; sudo apt-get install git
```

6. Clone the git directory into "`$mySoftware`":

```
git clone https://github.com/lh3/bwa.git
```

7. Change into the new "`bwa`" directory:

```
cd bwa
```

8. Make the file that will install BWA:

```
make
```

Install SAMtools

9. Change back into the "$mySoftware" directory.

```
cd $mySoftware
```

10. To install SAMtools, first download the current version into your $mySoftware directory. This will install version 1.9, current as of this writing:

```
wget https://github.com/samtools/samtools/releases/
download/1.9/samtools-1.9.tar.bz2
```

11. Next, extract the file with the "tar" command:

```
tar -xvjf samtools-1.9.tar.bz2
```

12. Change into the samtools directory:

```
cd samtools-1.9/
```

13. Then, change a configuration file to point to $mySoftware/samtools:

```
./configure --prefix=$mySoftware/samtools
```

14. Finally, build the SAMtools installation:

```
make; make install
```

Pull GATK and VEP images

15. Pull the GATK version 4.0.11.0 Docker image:

```
docker pull broadinstitute/gatk:4.0.11.0
```

16. Pull VEP v 95.1:

```
docker pull ensemblorg/ensembl-vep:release_95.1
```

ALTERNATE PROTOCOL

MUTATION DETECTION WITH MULTIPLE BIOLOGICAL REPLICATES

In practice, mutation detection is often performed on many mice to capture natural biological variation, to account for the variable penetrance of mutations, and to ensure statistical power to find variations of interest. Below, Alternate Protocol duplicates the

steps in Basic Protocol 2, but for multiple files. Of note, the entire process in Basic Protocol 1 (above) must be run for each file.

Materials

Hardware

Workstation (Mac OS X, Linux, or any Unix-like system) with 500 GB of RAM, a 64-bit processor with at least two real or virtual cores, and 500 GB free disk space either on the workstation or in an external drive. An alternative is a remote server meeting at least these specifications. Administrator privileges are required for initial software setup, but do not have to belong to the end user. See Computational Resource Requirements for more information.

Software

A JRE can be verified by visiting *https://www.java.com/en/download/installed.jsp?detect=jre*
GATK version 4.0.11.0
Docker

Datasets

For each input pair of FASTQ files from paired-end experiments, there should be a resulting deduplicated, sorted, indexed SAM file in the associated `$myProject` directory. When performed on one sample, the resulting sample files, and index files. The directory `musRefs_10` should contain everything generated in Basic Protocol 1:

```
$myProject/musRefs_10/ref_mm10.fasta
$myProject/musRefs_10/ref_mm10.dict
$myProject/musRefs_10/ref_mm10.fasta.fai
```

To proceed, assume that multiple samples have been processed generating each of the files below, with names "`sample1`" and "`sample2`" indicating the indexed BAM files for each of two replicates:

```
$myProject/sample1.sorted.rmDup.bam
$myProject/sample1.sorted.rmDup.bam.bai
$myProject/sample2.sorted.rmDup.bam
$myProject/sample2.sorted.rmDup.bam.bai
```

NOTE: See Basic Protocol 2 for a full explanation of the options passed to GATK below.

1. Within the `$myProject` directory, enter the GATK Docker container as in Basic Protocol 1, changing into the `$myProject` directory to batch process files:

```
docker run -t -i -v $PWD:/gatk/MouseMutations broad-
institute/gatk:4.0.11.0 /bin/bash
myProject=/gatk/MouseMutations
cd $myProject
```

2. In the code below, a variable "`f`" is created to hold the names of each sample BAM file, and an additional variable is created to hold a new name for each output file "`newF`." In a "for loop," call the GATK Haplotype Caller for each of the sample files above, generating a "`g.vcf`" file for each sample:

```
for f in *.sorted.rmDup.bam; do newF=${f/bam/g.vcf};
/gatk/gatk --java-options "-Xmx4G" HaplotypeCaller -R
$myProject/musRefs_10/ref_mm10.fasta -I $myProject/$f
-O $myProject/$newF -ERC GVCF; done;
```

This produces several `.g.vcf` files. These "`.g.vcf`" files produced by GATK need to be consolidated with the CombineGCVFs function, which will produce one `.g.vcf` for combined genotype calling. To enable this, first create a file of arguments to read into the CombineGCVFs function. The function takes a list of each variant on the command line, and this can be automated by creating a file listing these commands.

3. Below, search for each `.g.vcf` file and store it in the "`i`" variable, then paste each variable into a "`myArguments.txt`" file:

```
for i in *.g.vcf; echo "--variant $i" >> myArguments.
txt; done;
```

Next, use the list of annotations to combine the "`.g.vcf`" files with the Combine GVCFs function. Note that unlike the above analogous step in Basic Protocol 2, no "`-V`" flag is passed containing the "`.g.vcf`" file; instead, these files are listed in the "`myArguments.tt`" file.

4. Type the following to produce one `toFilter-gatk.vcf` file:

```
/gatk/gatk GenotypeGVCFs -R $myProject/musRefs_10/
ref_mm10.fasta -O $myProject/toFilter-gatk.vcf
--arguments_file myArguments.txt
```

This final step duplicates the last step of Basic Protocol 2, above (refer to Basic Protocol 2, step 3).

5. Type:

```
/gatk/gatk VariantFiltration -R $myProject/musRefs
_10/ref_mm10.fasta -V $myProject/toFilter-gatk.vcf
-O $myProject/filtered-gatk.vcf --filter-expression
"QUAL > 30.0 && DP == 10" --filter-name "myDepth
QualityFilter"
```

6. Lastly, exit the Docker container:

```
exit
```

SUPPORT PROTOCOL 2

VARIANT GENOTYPING WITH ADDITIONAL HARD FILTERING

As a substitute for the last step of Basic Protocol 2 and the identical step in Alternate Protocol, many additional filtering options may be passed to GATK's VariantFiltration function. Above, filtering was performed to filter on Phred score (base calling quality) and read depth. This support protocol demonstrates how to edit the command above to include other recommended filtering steps.

Materials

Hardware

Workstation (Mac OS X, Linux, or any Unix-like system) with 500 GB of RAM, a 64-bit processor with at least two real or virtual cores, and 500 GB free disk space either on the workstation or in an external drive. An alternative is a remote server meeting at least these specifications. Administrator privileges are required for initial software setup, but do not have to belong to the end user. See Computational Resource Requirements for more information.

Software

A JRE can be verified by visiting *https://www.java.com/en/download/installed.jsp?detect=jre*
GATK version 4.0.11.0
Docker

Datasets

The reference genome and associated files should be included as above in Basic Protocol 2:

```
$myProject/musRefs_10/ref_mm10.fasta
$myProject/musRefs_10/ref_mm10.dict
$myProject/musRefs_10/ref_mm10.fasta.fai
```

Input to the filtering step is the VCF file generated above:

```
$myProject/toFilter-gatk.vcf
```

1. Start the GATK Docker container as described above in Basic Protocol 1:

```
docker run -t -i -v $PWD:/gatk/MouseMutations
broadinstitute/gatk:4.0.11.0 /bin/bash
myProject=/gatk/MouseMutations
```

2. Apply the code below. The output file, "`filtered-gatk.vcf`," will be identical to that at the end of Basic Protocol 2, except different variants will have a "`PASS`" flag, and the new filter name is "`myComplexFilter`." Type:

```
/gatk/gatk VariantFiltration -R $myProject/
musRefs_10/ref_mm10.fasta -V $myProject/toFilter-
gatk.vcf -O $myProject/filtered-gatk.vcf --filter-
expression "QD < 2.0 || FS > 60.0 || MQ < 40.0
|| MQRankSum < -12.5 || ReadPosRankSum < -8.0"
--filter-name "myComplexFilter"
```

Each flag above performs a QC step. `QD < 2` divides the variant confidence (QUAL) by unfiltered read depth of all model samples. `MQ < 40.0` accesses mapping quality across all samples. `FS > 60` performs a Fisher's exact test to detect strand bias. `MQRankSum < --13.5` tests differences in mapping qualities between reference and alternate alleles using a z-approximation of a Mann-Whitney U-test. `ReadPosRankSum < --8.0` applies this test to the distance of an alternative allele to the end of a read to screen false positives in heterozygous alleles.

The "||" symbol is a logical OR, indicating that a `myComplexFilter` flag will be entered for a variant if any of the above conditions are met. Additional information can be found via GATK documentation: *https://software.broadinstitute.org/gatk/documentation/article?id=11069*.

3. Exit the GATK Docker container:

```
exit
```

COMMENTARY

Background Information

Mutation detection begins with the alignment of NGS reads to the mouse reference genome (see Fig. 1). Although there are over three billion bases in the mouse genome, reads of length ≥25 are typically sufficient to cover the entire mouse genome. Short single-end reads are at a disadvantage, as they may contain a run of bases that are homologous to various regions of the genome, and as such can be placed at various regions of the genome, producing erroneous mapped regions that may have a downstream erroneous effect. The more popular paired-end reads contain a central spacer of approximately 125 bases and are mapped to positions where both reads are placed together. This placement reduces the problem of numerous mapping events due to homologous regions and increases the number of unique mapping events even when there are mismatches or gaps (Chen et al., 2012). The number of reads that map to a given genomic location is termed read depth or coverage. Sequence variants in the aligned genome are deviations from the reference sequence; these range from small differences including SNPs and indels to the larger and complicated variations including inversions and translocations. In a sequence alignment, heterozygous variations manifest as positions where approximately half of the reads match the reference and the other reads differ from the reference, although the theory rarely matches reality.

There has been a lot of debate on how much read depth or coverage is required to successfully call a small sequence variant such as a SNP or small indel. A sequencing run performed by one of the large sequencing companies such as Illumina (Goodwin, McPherson, & McCombie, 2016) generates reads randomly, but they are not distributed equally across the genome. Instead, some regions of the genome are better covered by sequence reads than others (Sims, Sudbery, Ilott, Heger, & Ponting, 2014). For instance, repetitive DNA regions are poorly covered because they can generate multiple, very similar short sequence reads, and the mapping of these regions may puzzle the assembly program, resulting in no or ambiguous mapping and erroneous alignments. Thus, the more frequently a base is sequenced, the more reliable that base is. Sequence variant detection with <20× read depth is commonplace (Potter et al., 2016), as is multiplexing samples with a shallow read depth (Bull et al., 2013). However, for detecting homozygous SNPs, it is advisable to have a read depth of 15×, and heterozygous SNPs should have a read depth of 33× (Bentley et al., 2008). Alongside the read depth are other parameters that are important to mutation detection, including base-mapping quality and the mapping quality of the surrounding sequence. Many SNP callers have built-in scoring systems to distinguish a genuine SNP from an error. For example, when GATK encounters a region of high variation, it reassembles mapped reads in that region, and then uses a hidden Markov model to determine the likelihood of haplotypes (and then alleles) for each variant site. By default, filters are included based on mapping quality and read quality, specifically according to the Phred score of each called base and read depth. Filtering by read depth ensures that a sufficient number of reads were aligned to each base. Phred scores come precalculated by sequencing machines based on a standard reference recovery rate for the specific chemistry of any given sequencer. These scores are encoded in FASTQ files directly, and allow filtering by the log likelihood of miscalling any given base. Furthermore, filtration programs (see Support Protocol 2) filter for strand bias, mapping quality, and sequencing artifacts, including combination scores such as QD (quality by depth), which incorporates a Phred-scaled probability that samples are homozygous for a reference allele and read depth.

When calling variants, two critical choices may affect results: the choice of the transcript set to use (Ensembl, RefSeq, UCSC) and the choice of variant calling software (SAMtools or GATK, for example). As many bioinformatics tools accept Ensembl transcript IDs as input for further analysis, and McCarthy et al. (2014) found Ensembl to outperform Refseq when discovering loss of function variants, we recommend using Ensembl transcripts and have done so in this article. GATK's realignment of variant-rich regions produces a higher positive predictive value than SAMtools; thus, it is recommended for variant detection in mouse, especially when multiple variants may be expected (Pirooznia et al., 2014). There are many variant annotators freely available to annotate SNPs. Among the most popular are VEP and Annovar (Wang, Li, & Hakonarson, 2010). Discrepancies in definitions of sequence variants partially explain the difference in annotation recovery between VEP and Annovar (McCarthy et al., 2014). VEP currently annotates sequences using, among other tools, the Sequence Ontology, which allows the specific variant types used by VEP to be consistent across experiments and databases (Eilbeck et al., 2005). Lastly, VEP was chosen because of the Docker implementation. Docker enables software to be managed easily and updated within a standard environment; this will facilitate expanding this article as newer variant detection algorithms become available in the future.

It is important to note that this protocol is optimized for detecting germline mutations. When detecting and annotating somatic (cancer) mutations, different tools should be used to take into account extreme aneuploidy and incorporate a "panel of normals." Such a panel is taken from normal (noncancerous) tissue sequenced with the same technology as the mutant data. It seeks to capture technical artifacts and thus improve the accuracy of variant calling.

Troubleshooting

Troubleshooting is summarized in Table 4. For additional problems, each tool installed (BWA, FASTQC, SAMTools, GATK, VEP) has documentation and can be run with "`--help`" flags.

The Java Virtual Machine must be assigned memory to handle large datasets encountered in mouse genome experiments. If, when executing any GATK, Picard Tools, or FASTQC programs an error occurs, which references an out of memory error, for example:

```
Exception in thread "main"
java.lang.OutOfMemoryError:
Java heap space
```

then an additional argument must be made to each call to the program in question. Instead of typing:

```
/gatk/gatk HaplotypeCaller
... {other options}
```

type:

```
/gatk/gatk --java-options
"-Xmx4G" HaplotypeCaller
... {other options}
```

The "`-Xmx4G`" option allocates 4 GB of RAM to the Java Virtual Machine. This option can be increased to fit the workstation memory limits and the size of the datasets being analyzed.

GATK is called with a wrapper script. To fix this issue with any other program directly called by the Java program itself, use the following syntax:

```
java -jar program.jar
[program arguments]
```

Understanding Results

The anticipated results are ultimately reports from VEP, as described in Basic Protocol 4, yielding both a text file of raw results and an html file of pre-made summary statistics. SNVs and structural variations may be identified. Within the text file of annotations, variations will be annotated with known rsIDs, Ensembl gene IDs for any intergenic variations found, and possible consequences of each variation. This may later be mined for more information relating to variants themselves, or be combined with previous knowledge to inform systems biology studies.

Time Considerations

Each protocol varies in time of execution. Once Support Protocol 1 (installation of software) is completed, Basic Protocol 1 is the largest consumer of time. Indexing the mouse genome takes several hours on one processor. By multithreading in a server environment, the alignment steps in Basic Protocol 1 can be reduced from days (when many replicates are involved) to hours. Basic Protocol 2 and Support Protocol 2 likewise take several hours; with multithreading via Spark, GATK

Table 4 Troubleshooting Guide for Mutation Detection in Mouse[a]

Problem	Possible cause	Solution
Basic Protocol 4 produces multiple results	VEP was run on a multisample VCF	Use GATK to combine each VCF before running Basic Protocol 4, or run as separate processes
No variants flagged as "`PASS`" in GATK	Low coverage	Pool samples of low coverage together if appropriate via the HaplotypeCaller and GenotypeGVCF
Command not found	Typos	Re-check typing in command prompt
Java heap space error	Low Java memory	Run Java command with "`Xmx4G`" flag
Java heap space error	Low Docker container memory	Run Docker with:`docker run -m=4g {imageID}` where 4g is 4 GB; increase this as needed
Files not found	`$myProject/ $mySoftware` not initialized	Re-initialize bash variables as in Basic Protocol 1
Error response from daemon; bad response from Docker engine	Docker engine not started	Start Docker. Mac OS: run open `--background -a Docker`. Linux: `sudo systemctl start docker`

[a]For help with specific problems not covered, check the documentation for each program. Always make sure to check for typographical errors, and be sure that each command is executed in the proper working directory (folder). Lastly, make sure that any bash variables such as "`myProject`" are initialized.

filtering steps can be completed in 2 h for small samples. As software is constantly improving, speeds are expected to greatly improve with increased multithreading support. Time of execution of Basic Protocols 3 and 4 largely depend on the number of novel and previously known variants found, but it is expected that this step will be completed in a matter of hours. Interpretation of variants and identifying variants with biological meaning in the context of any given experiment may take time depending on a researchers' domain expertise.

Acknowledgments

Research reported in this publication was supported by the Medical Research Council (MC_U142684171) and National Human Genome Research Institute of the National Institutes of Health (UM1HG006370). The authors also would like to Gareth Banks for help on the breeding scheme.

Conflicts of Interest

The authors declare no conflicts of interest. The funding organizations had no role in the design of this study, data collection, analysis or interpretation, or preparation of the manuscript, and did not approve, disapprove, or delay publication of the work.

Literature Cited

Adams, D. J., Doran, A. G., Lilue, J., & Keane, T. M. (2015). The Mouse Genomes Project: A repository of inbred laboratory mouse strain genomes. *Mammalian Genome: Official Journal of the International Mammalian Genome Society*, *26*(9–10), 403–412. doi: 10.1007/s00335-015-9579-6.

Andews, S. (2015). FastQC: A quality control tool for high throughput sequence data (version 0.11.4). Retrieved from *https://www.bioinformatics.babraham.ac.uk/projects/fastqc/*.

Balling, R. (2001). ENU mutagenesis: Analyzing gene function in mice. *Annual Review of Genomics and Human Genetics*, *2*(1), 463–492. doi: 10.1146/annurev.genom.2.1.463.

Bennett, S. (2004). Solexa Ltd. *Pharmacogenomics*, *5*(4), 433–438. doi: 10.1517/14622416.5.4.433.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., ... Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59. doi: 10.1038/nature07517.

Bull, K. R., Rimmer, A. J., Siggs, O. M., Miosge, L. A., Roots, C. M., Enders, A., ... Cornall, R. J. (2013). Unlocking the bottleneck in forward genetics using whole-genome sequencing and identity by descent to isolate causative mutations. *PLOS Genetics*, *9*(1), e1003219. doi: 10.1371/journal.pgen.1003219.

Chen, Y., Negre, N., Li, Q., Mieczkowska, J. O., Slattery, M., Liu, T., ... Liu, X. S. (2012). Systematic evaluation of factors influencing ChIP-seq fidelity. *Nature Methods*, *9*(6), 609–614. doi: 10.1038/nmeth.1985.

Concepcion, D., Seburn, K. L., Wen, G., Frankel, W. N., & Hamilton, B. A. (2004). Mutation rate and predicted phenotypic target sizes in ethylnitrosourea-treated mice. *Genetics*, *168*(2), 953–959. doi: 10.1534/genetics.104.029843.

Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., & Ashburner, M. (2005). The Sequence Ontology: A tool for the unification of genome annotations. *Genome Biology*, *6*(5), R44. doi: 10.1186/gb-2005-6-5-r44.

Fabbro, C. D., Scalabrin, S., Morgante, M., & Giorgi, F. M. (2013). An extensive evaluation of read trimming effects on illumina NGS data analysis. *PLoS One*, *8*(12), e85024. doi: 10.1371/journal.pone.0085024.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333–351. doi: 10.1038/nrg.2016.49.

Justice, M. J., Noveroske, J. K., Weber, J. S., Zheng, B., & Bradley, A. (1999). Mouse ENU mutagenesis. *Human Molecular Genetics*, *8*(10), 1955–1963. doi: 10.1093/hmg/8.10.1955.

Korlach, J., Gedman, G., Kingan, S. B., Chin, C.-S., Howard, J. T., Audet, J.-N., ... Jarvis, E. D. (2017). De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. *GigaScience*, *6*(10), 1–16. doi: 10.1093/gigascience/gix085.

Krueger, F. (2019). Trim Galore (version 0.6.0). Retrieved from *https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/*.

Leinonen, R., Sugawara, H., & Shumway, M., & International Nucleotide Sequence Database Collaboration. (2011). The sequence read archive. *Nucleic Acids Research*, *39*(Database issue), D19–D21. doi: 10.1093/nar/gkq1019.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. doi: 10.1093/bioinformatics/btp324.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., & Homer, N. ... 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. doi: 10.1093/bioinformatics/btp352.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10–12. doi: 10.14806/ej.17.1.200.

McCarthy, D. J., Humburg, P., Kanapin, A., Rivas, M. A., Gaulton, K., & Cazier, J.-B. ... The WGS500 Consortium. (2014). Choice of transcripts and software has a large effect on variant annotation. *Genome Medicine*, *6*(3), 26. doi: 10.1186/gm543.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. doi: 10.1101/gr.107524.110.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., ... Cunningham, F. (2016). The Ensembl variant effect predictor. *Genome Biology*, *17*(1), 122. doi: 10.1186/s13059-016-0974-4.

Merkel, D. (2014). Docker: Lightweight Linux containers for consistent development and deployment. *Linux Journal*, *2014*(239). Retrieved from *http://dl.acm.org/citation.cfm?id=2600239.2600241*.

Mianné, J., Chessum, L., Kumar, S., Aguilar, C., Codner, G., Hutchison, M., ... Bowl, M. R. (2016). Correction of the auditory phenotype in C57BL/6N mice via CRISPR/Cas9-mediated homology directed repair. *Genome Medicine*, *8*(1), 16.

Mouse Genome Sequencing Consortium. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, *420*(6915), 520–562. doi: 10.1038/nature01262.

O'Brien, T. P., & Frankel, W. N. (2004). Moving forward with chemical mutagenesis in the mouse. *The Journal of Physiology*, *554*(1), 13–21. doi: 10.1113/jphysiol.2003.049494.

Panda, S. K., Wefers, B., Ortiz, O., Floss, T., Schmid, B., Haass, C., ... Kühn, R. (2013). Highly efficient targeted mutagenesis in mice using TALENs. *Genetics*, *195*(3), 703–713. doi: 10.1534/genetics.113.156570.

Park, S. T., & Kim, J. (2016). Trends in next-generation sequencing and a new Era for

whole genome sequencing. *International Neurology Journal*, *20*(Suppl 2), S76–S83. doi: 10.5213/inj.1632742.371.

Picard Toolkit. (2018). Retrieved from *http://broadinstitute.github.io/picard/*.

Pirooznia, M., Kramer, M., Parla, J., Goes, F. S., Potash, J. B., McCombie, W. R., & Zandi, P. P. (2014). Validation and assessment of variant calling pipelines for next-generation sequencing. *Human Genomics*, *8*, 14. doi: 10.1186/1479-7364-8-14.

Potter, P. K., Bowl, M. R., Jeyarajan, P., Wisby, L., Blease, A., Goldsworthy, M. E., . . . Brown, S. D. M. (2016). Novel gene function revealed by mouse mutagenesis screens for models of age-related disease. *Nature Communications*, *7*, 12444. doi: 10.1038/ncomms12444.

Qian, L., Mahaffey, J. P., Alcorn, H. L., & Anderson, K. V. (2011). Tissue-specific roles of Axin2 in the inhibition and activation of Wnt signaling in the mouse embryo. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(21), 8692–8697. doi: 10.1073/pnas.1100328108.

Sengupta, S., Bolin, J. M., Ruotti, V., Nguyen, B. K., Thomson, J. A., Elwell, A. L., & Stewart, R. (2011). Single read and paired end mRNA-Seq Illumina libraries from 10 nanograms total RNA. *Journal of Visualized Experiments*, *56*, e3340. doi: 10.3791/3340.

Shendure, J. A., Porreca, G. J., & Church, G. M. (2008). Overview of DNA sequencing strategies. *Current Protocols in Molecular Biology*, *81*, 7.1.1–7.1.11. doi: 10.1002/0471142727.mb0701s81.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: Key considerations in genomic analyses. *Nature Reviews Genetics*, *15*(2), 121–132. doi: 10.1038/nrg3642.

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164. doi: 10.1093/nar/gkq603.

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster computing with working sets. *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*, June 22–25, Boston, MA. Retrieved from *http://dl.acm.org/citation.cfm?id=1863103.1863113*.

Key References

Butkiewicz, M., & Bush, W. S. (2016). In silico functional annotation of genomic variation. *Current Protocols in Human Genetics*, *88*(1), 6.15.1–6.15.17. doi: 10.1002/0471142905.hg0615s88.

Buteiwcz and Bush (2016) provide a detailed overview variant annotation, explaining why certain elements may be annotated for biological importance by means of the Sequence Ontology.

Internet Resources

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

For information on evaluating the QC status of your reads, the FastQC website has excellent tutorials outside the scope of this protocol. As of this writing date, the following tutorial video is available from the Babraham institute via YouTube: http://www.youtube.com/watch?v=bz93ReOv87Y.

https://www.ensembl.org/info/docs/tools/vep/index.html.

See the link above for access to VEP's web interface, which provides an alternative means of uploading a VCF file and retrieving results. For direct links to sources which VEP uses for annotating mouse variants, including a version number for each resource, see: https://www.ensembl.org/info/genome/variation/species/sources_documentation.html#mus_musculus.

https://software.broadinstitute.org/gatk/documentation/.

The user guide for GATK provides best practices, links to tool documentation, and information about the continual growth of the GATK resource.