I. Streszczenie

Klasyczne geny głównego układ zgodności tkankowej (ang. *major histocompatibility complex*, MHC) kodują cząsteczki prezentujące antygeny limfocytom T - komórkom odpowiadającym za swoistą odpowiedź odpornościową kręgowców (Klein 1986). Wyróżnia się dwa typy klasycznych cząsteczek MHC: klasę I i II, które różnią się budową, dystrybucją tkankową oraz funkcją. MHC klasy I zbudowane są z ciężkiego łańcucha α oraz lekkiego mβ2, występują na powierzchni wszystkich jądrzastych komórek i są odpowiedzialne za prezentację antygenów pochodzenia wewnątrzkomórkowego (np. wirusów lub bakterii namnażających się w cytoplazmie) limfocytom cytotoksycznym. MHC klasy II budują podobne do siebie łańcuchy α i β, występują one głównie na wyspecjalizowanych komórkach prezentujących antygen (np. dendrytycznych, limfocytach B, makrofagach) i są odpowiedzialne za prezentację antygenów pochodzenia zewnątrzkomórkowego (np. większość bakterii, makropasożyty) limfocytom pomocniczym (Klein 1986; Kindt et al. 2007).

O ile MHC jest konieczne do wiązania antygenów, reakcja układu odpornościowego pochodzi od limfocytów T typu αβ. Jest ona inicjowana poprzez swoiste rozpoznanie antygenu związanego z MHC, przez receptor limfocytu T (ang. *T-cell receptor*, TCR). TCR to heterodimeryczne cząsteczki zbudowane łańcuchów α i β. Każdy łańcuch posiada część stałą i zmienną, w przypadku łańcucha α część zmienna jest kodowana przez segmenty genowe V (od ang. *variable*) i J (ang. *joining*); część zmienną łańcucha β tworzą segmenty V, D (ang. *diversity*) i J. Olbrzymia zmienność receptorów jest generowana, podobnie do przeciwciał, na drodze somatycznej rekombinacji w/w segmentów genowych. Poza kombinacyjną zmiennością wynikającą z możliwości łącznia wielu obecnych w genomie segmentów V, D i J, losowe nukleotydy są wstawiane bądź usuwanie na złączach pomiędzy segmentami, tworząc tzw. regiony N. Znajdująca się na łączeniach segmentów V(D)J hiper-zmienna część łańcucha koduje tzw. trzeci region determinujący dopasowanie: CDR3 (ang. *complementarity-determining region*). To właśnie ten region bierze udział w swoistym rozpoznaniu antygenu prezentowanego przez MHC (Davis and Bjorkman 1988; Kindt et al. 2007).

W szerszym kontekście, interakcja pomiędzy antygenami prezentowanymi przez MHC, a limfocytami T, jest podstawą molekularnego rozróżnienia własnych cząsteczek od obcych. W zdrowym, prawidłowo funkcjonującym organizmie, rozpoznanie obcego antygenu powoduje rozpoczęcie swoistej odpowiedzi odpornościowej (Kindt et al. 2007). Własne cząsteczki są zaś ignorowane, dzięki nabytej podczas dojrzewania w grasicy tzw. tolerancji centralnej. Odpowiedzialne są za nią dwa główne mechanizmy: pozytywna i negatywna selekcja (Klein et al. 2009, 2014). W trakcie pozytywnej selekcji, limfocyty T rozpoznające własne antygeny na powierzchni MHC, otrzymują sygnał pozwalający im na przeżycie. Brak takiej interakcji prowadzi do śmierci z zaniedbania. W trakcie negatywnej selekcji, usuwane są zaś te limfocyty, które zbyt silne rozpoznają kompleks MHC-własny antygen, gdyż mogłyby one powodować odpowiedź autoimmunologiczną (Klein et al. 2014). Biorąc pod uwagę element losowości obecny w trakcie tworzenia receptorów limfocytów T, znaczna ich większość jest niefunkcjonalna lub auto-reaktywna, i w efekcie zostaje usunięta w trakcie dojrzewania w grasicy. Szacuje się, że dojrzałość osiąga jedynie 5% limfocytów T, przy czym 20-25% przechodzi przez pozytywną selekcję, a następnie 20-50% z nich przeżywa selekcję negatywną (przegląd w Yates 2014). Tak powstaje różnorodny i funkcjonalny repertuar TCR dojrzałych limfocytów T, który jest zdolny do rozpoznawania obcych antygenów w kontekście obecnych u danego osobnika alleli MHC, ale tolerancyjny wobec własnych tkanek.

Zasadniczym tematem mojej rozprawy doktorskiej jest hipoteza zakładająca, że proces negatywnej selekcji w grasicy może przyczyniać się do ewolucyjnego kompromisu limitującego wewnątrz-osobniczą ekspansję i różnicowanie klasycznych genów MHC. Ze względu na swój olbrzymi polimorfizm, geny MHC od lat są obiektem badań biologii ewolucyjnej. W populacjach zwykle opisuje się dziesiątki lub setki alleli, a w niektórych przypadkach (np. u ludzi) są to nawet tysiące allelicznych wariantów (Apanius et al. 1997; Bernatchez and Landry 2003). Za źródło tak dużej różnorodności uważa się ewolucyjny „wyścig zbrojeń” pomiędzy patogenami a gospodarzami (Jeffery and Bangham 2000; Spurgin and Richardson 2010). Różnice pomiędzy poszczególnymi allelami kumulują się we fragmentach wiążących peptydy; obserwuje się tam podwyższone tempo podstawień niesynonimicznych do synonimicznych (Reche and Reinherz 2003). Jest to interpretowane jako działanie doboru pozytywnego, który różnicując sekwencje aminokwasowe MHC sprawia, że poszczególne allele są w stanie

wiązać odmienne rodzaje antygenów. Następnie, tak duża zmienność jest utrzymywana w populacjach przez różne formy do doboru równoważącego, takie jak przewaga heterozygot (większe dostosowanie osobników heterozygotycznych wynikające z możliwości rozpoznania większego spektrum patogenów; Doherty and Zinkernagel 1975; Penn et al. 2002; Oliver et al. 2009), czy dobór negatywnie zależny od częstości (większe dostosowanie posiadaczy rzadkich lub nowych alleli MHC, ze względu na adaptacje patogenów do „unikania" wiązania przez częste warianty MHC; Borghans et al. 2004; Ejsmond and Radwan 2015; Phillips et al. 2018).

Wiele dekad badań ewolucyjnych poświęcono pochodzeniu i mechanizmom utrzymującym olbrzymi polimorfizm MHC w populacjach (Bernatchez and Landry 2003; Spurgin and Richardson 2010), jednak mechanizmy kształtujące liczbę genów MHC na poziomie osobniczym są znacznie gorzej poznane. Biorąc pod uwagę adaptacyjną wartość polimorfizmu MHC, przewagę heterozygot oraz ewolucyjny potencjał do duplikacji genów – zaskakująca wydaje się niewielka liczba funkcjonalnych loci na poziomie osobniczym. Co prawda zwykle jest ich przynajmniej kilka, jednak poszczególnym osobnikom pozwala to na ekspresję jedynie niewielkiej części różnorodności obecnej na poziomie populacji. Wytłumaczenie tego paradoksu zostało zaproponowane ponad 30 lat temu (Vidović and Matzinger 1988), i później sformalizowane na gruncie matematycznym jako *hipoteza optymalności* (Nowak et al. 1992). Przewiduje ona kompromis ewolucyjny pomiędzy możliwością rozpoznania większego spektrum patogenów (związaną z ekspansją MHC), a zmniejszeniem różnorodności limfocytów T (związanych z działaniem mechanizmów zapewniających tolerancję centralną). Wiązanie większej różnorodności peptydów dotyczyłoby także własnych antygenów, więc większa liczba receptorów TCR mogłaby okazać się auto-reaktywna. Pociągałoby to za sobą znaczące zmniejszenie dostępnego repertuaru TCR, wynikające z nasilenia negatywnej selekcji grasiczej - stąd alternatywna nazwa: *hipoteza deplecji TCR* (Woelfing et al. 2009). Krytycy tej hipotezy zauważyli jednak, że zwiększenie możliwości prezentacji własnych antygenów powinno, w pierwszej kolejności, zwiększyć również pulę limfocytów selekcjonowanym pozytywnie (Borghans et al. 2003). W zależności od przyjętych parametrów, teoria przewiduje więc zmniejszenie (Nowak et al. 1992; Woelfing et al. 2009) lub względne zwiększenie (Borghans et al. 2003) repertuaru TCR wraz ze wzrostem liczby i różnorodności genów MHC. Rozstrzygnięcie, który z modeli lepiej opisuje rzeczywiste efekty edukacji grasiczej utrudnia fakt, że kluczowe dla nich parametry i założenia są wciąż tematem

żywych dyskusji pomiędzy immunologami (Yates 2014; La Gruta et al. 2018). W szczególności, wciąż nie jest pewne, jak dużą rolę w pozytywnej selekcji odkrywają mniej zmienne i ewolucyjnie zakonserwowane części MHC i TCR, w porównaniu do roli poszczególnych antygenów wewnątrz polimorficznej kieszeni MHC i rozpoznających je hiper-zmiennych regionów CDR3 (Scott-Browne et al. 2009; Holland et al. 2012; Baker and Evavold 2017).

Ekolodzy i biolodzy ewolucyjni w odmienny sposób podeszli to kwestii immunogenetycznej optymalności, i starali się dostarczyć pośrednich dowodów na słuszność tej hipotezy. W swoich badaniach wykorzystywali gatunki, u których występują między-osobnicze różnice w liczbie genów MHC. Fenomen ten często nazywany jest zmienną liczbą kopii genów (ang. *copy number variation*, CNV), chociaż nie dotyczy identycznych kopii, lecz raczej zduplikowanych loci, które uległy dywersyfikacji (Siddle et al. 2010; Lighten et al. 2014). Zakładano, że osobniki posiadające pośrednią - „optymalną" - liczbę genów MHC, będą miały najwyższą immunokompetencję lub dostosowanie. Badania korelujące indywidualną liczbą wariantów MHC z intensywnością lub podatnością na infekcje pasożytnicze, całożyciowym sukcesem reprodukcyjnym, lub innymi wskaźnikami dostosowania (rozmiar ciała, depozyty tkanki tłuszczowej) osiągały jednak niespójne, a często sprzeczne wyniki (Wegner et al. 2003; Bonneaud et al. 2004; Madsen and Ujvardi 2006; Radwan et al. 2012; Westerdahl et al. 2012; Sepil et al. 2013; Biedrzycka et al. 2018). Wynikać to może z pośredniego charakteru tych badań, gdyż na korelację pomiędzy liczbą wariantów MHC, a statusem/intensywnością infekcji lub innymi komponentami dostosowania, może wpływać wiele czynników, jak chociażby efekt konkretnych alleli lub zmienna presja pasożytnicza w czasie i przestrzeni (O'Connor et al. 2018).

Mimo znaczenia hipotezy optymalności dla zrozumienia ewolucji układu odpornościowego kręgowców, oraz jej implikacji sięgających obszarów tak odległych od immunogenetyki ewolucyjnej jak np. teorie mechanizmów specjacji (Eizaguirre et al. 2009; Malmstrøm et al. 2016), przez długi czas brakowało rozstrzygającego, bezpośredniego testu jej przewidywań. W szczególności, nie wykazano negatywnej zależności pomiędzy osobniczą, wysoką liczbą genów MHC, a zmniejszeniem repertuaru TCR. Wynikało to głównie z technicznych trudności związanych z sekwencjonowaniem i analizą olbrzymich repertuarów immunologicznych, takich jak te receptorów limfocytów. Zmieniło się to z nastaniem ery wysoko przepustowego

sekwencjonowanie nowej generacji (HTS, ang. *high throughput sequencing*) (Benichou et al. 2012).

Głównym celem mojej pracy doktorskiej było **testowanie kluczowych przewidywań *hipotezy optymalności*: sprawdzenie, czy osobniki posiadające większą liczbę wariantów genów MHC charakteryzują się zmniejszonym repertuarem TCR**. Obiektem badań była nornica ruda (*Myodes glareolus*), niewielki gryzoń z rodziny chomikowatych, często wykorzystywany jako model w badaniach ekologicznych i ewolucyjnych. Na potrzeby pracy wykorzystano tkanki nornic pochodzących z laboratoryjnej hodowli Instytutu Nauk o Środowisku UJ (dzięki uprzejmości prof. Pawła Kotei). Gatunek ten został wybrany głownie ze względu na wysoką, międzyosobniczą zmienność w liczbie kopii genów MHC (wcześniej udokumentowano taką zmienność dla MHC klasy II, np. Axtner & Sommer, 2007; Bryja et al., 2006). Jednakowoż, jako, że nie jest to typowy gatunek modelowy biologii molekularnej, istniały poważne luki w dostępnych dla niego zasobach immunogenetycznych. Dlatego też, aby zrealizować zamierzony cel badawczy, niezbędna była najpierw **charakterystyka kluczowych genów zaangażowanych w prezentację i rozpoznanie patogenów u nornicy rudej**, czyli opis zmienności na poziomie MHC (w szczególności wcześniej nieopisywanej u tego gatunku MHC klasy I, oraz zmienności w liczbie ulegających ekspresji genów MHC klasy II), oraz wielkości repertuaru receptorów limfocytów T.

W pierwszej części mojej rozprawy scharakteryzowałam ulegające ekspresji geny MHC klasy I u nornicy rudej (Migalska et al. 2017)**.** Wcześniej dostępne były prace charakteryzujące poszczególne ortologii MHC klasy II u tego gatunku: DQA (Bryja et al. 2006; Deter et al. 2008), DQB (Scherman et al. 2014) oraz DRB (Axtner and Sommer 2007; Kloch et al. 2010)), jednak brakowało opisu MHC klasy I. Wykorzystałam *de novo* składane transkryptomy pochodzące ze śledzion 7 osobników, aby zaprojektować specyficzne dla tego gatunku startery. Następnie, otrzymałam region kodujący cząsteczki MHC klasy I (egzony 1-8 oraz część 3'UTR). Matryca cDNA została namnożona w reakcji PCR, klonowana w wektorze bakteryjnym oraz sekwencjonowana metodą Sangera. Wykorzystując egzony 2-4, kodujące zewnątrzkomórkowy fragment cząsteczki MHC, przeprowadziłam analizę filogenetyczną przyrównującą MHC klasy I u nornicy rudej do sekwencji gryzoni z rodziny chomikowatych oraz myszowatych. Geny nornic nie były ortologiczne do

genów pozostałych gatunków z rodziny chomikowatych, co jest typowe dla MHC klasy I – charakteryzującej się szybszym tempem ewolucji od MHC klasy II. Ponadto, do bardziej szczegółowych analiz polimorfizmu, u 29 osobników sekwencjonowałam metodą HTS egzon trzeci MHC klasy I. Do genotypowania MHC wykorzystałam powstały w Pracowni Biologii Ewolucyjnej (PBE) UAM program AmpliSAS (Sebastian et al. 2016). Już w tej niewielkiej grupie osobników obecna była duża zmienność w liczbie ulegających ekspresji genów: od 5 do 14 wariantów na osobnika. Ponadto, przeprowadziłam analizę sygnatur doboru, poprzez porównanie tempa podstawień synonimicznych do niesynonimicznych (d_N/d_S). Wykazała ona 8 kodonów znajdujących się pod doborem pozytywnym, z których większość pokrywa się z resztami zaangażowanymi w wiązanie peptydu u człowieka.

W drugiej część mojej rozprawy opisałam segmenty V i J budujące łańcuchy β receptorów TCR, scharakteryzowałam zmienność regionu CDR3 oraz estymowałam wielkość repertuaru TCRβ u nornicy rudej (Migalska et al. 2018). Cześć jakościowa tej pracy opisywała (na przykładzie 7 osobników) fragment zmienny łańcucha TCRβ, obejmującej zrekombinowane segmenty V, D i J. Do ich analizy zastosowałam metodę 5'RACE (*Rapid Amplification of cDNA Ends*), która wykorzystuje jeden starter umieszczony na końcu 5' regionu stałego, oraz uniwersalny adapter dołączany do końca 5' cDNA powstającego z transkryptu TCRβ (Mamedov et al. 2013). Startery dla tej reakcji zostały zaprojektowane na postawie *de novo* składanych transkryptomów, tak jak opisano w przypadku MHC klasy I. Tak uzyskana matryca była namnażana z wykorzystaniem odpowiednich, uniwersalnych starterów i sekwencjonowana z wykorzystaniem HTS na platformie Illumina MiSeq (długość odczytów 2×300 pz, co pozwala na odczyt całej długości tego fragmentu). Do analizy bioinformatycznej uzyskanych danych wykorzystałam stworzony w PBE UAM program AmpliTCR i AmpliCDR3 (Migalska et al. 2018). Zidentyfikowałam 37 grup genów V oraz 11 grup genów J – grupy te najprawdopodobniej odpowiadają loci. Analizy filogenetyczne wykazały ogólne zachowanie ortologii pomiędzy grupami loci nornic i myszy, mimo ogólnego zróżnicowania tych grup genów. Ponadto opisałam rozkład długości regionu CDR3 oraz scharakteryzowałam częstość parowania poszczególnych segmentów V i J u nornicy rudej – tego typu badania nie były do tej pory przeprowadzani u tego gatunku, oraz były pierwszym tego typu badaniem przeprowadzonym u organizmu niemodelowego. Część ilościowa tej pracy polegała na estymacji wielkości repertuaru TCRβ u nornicy rudej (na przykładzie 3 osobników).

Wykorzystałam zmodyfikowaną metodę molekularną 5'RACE, z adapterem dla końca 5' zawierającym unikalne identyfikatory molekularne (UMIs, ang. *unique molecular identifiers*) (Shugay et al. 2014). Wykorzystanie UMIs pozwala na efektywną identyfikację i eliminację błędów wprowadzanych na etapie PCR i sekwencjonowania, co jest szczególnie istotne przy analizie olbrzymiej zmienności repertuarów immunologicznych. Dla każdego osobnika wykonano 4 replikaty PCR, które następnie sekwencjonowane z wysokim pokryciem (>1 mln odczytów) na platformie HiSeq2500. Analizy bioinformatyczne zostały wykonane przy użyciu programu AmpliCDR3. Następnie, stosując estymator zaczerpnięty z badań nad bogactwem gatunkowym - Chao 2 (Chao 1987), - oszacowałam zakres wielkości repertuaru TCRβ u tego gatunku. W analizach wykorzystałam podzbiór 1,5 miliona odczytów na amplikon, aby uniknąć różnic wynikających z nierównej liczby odczytów. Średnia liczba unikalnych wariantów TCRβ CDR3 na osobnika (łącznie z 4 replikatów) wynosiła $1,8 \times 10^5$ sekwencji nukleotydowych, i $1,5 \times 10^5$ sekwencji aminokwasowych. Dolna granica estymowanej wielkości repertuary wynosiła średnio wynosiła $2,1 \times 10^5$ sekwencji nukleotydowych.

Trzecia część tej rozprawy prezentuje wyniki testu przewidywań hipotezy optymalności (Migalska et al. 2019). Na początku określiłam zmienność w liczbie ulegających ekspresji genów MHC klasy I i II w grupie ok. 150 osobników. Dla każdego markera (fragmentu genu namnażanego parą starterów) wykonałam dwie niezależne reakcje PCR, których produkty były sekwencjonowane instrumentem PGM Ion Torrent. Do genotypowania wykorzystałam metodę opracowaną w pierwszej części rozprawy (z pewnymi modyfikacjami). Te wstępne badania potwierdziły, że wysoka, międzyosobnicza zmienność w liczbie genów MHC klasy II jest obecna nie tylko na poziomie DNA u tego gatunku, ale również ulega ekspresji (DQA: 2-6, DQB: 2-8, DRB: 1-8). Następnie, wybrałam po około 40 osobników o niskiej i wysokiej liczbie wariantów MHC, i powtórzyłam u nich amplifikację i sekwencjonowanie MHC (tym razem na instrumencie Illumina MiSeq), uzyskując kolejne dwa replikaty PCR dla każdego markera. Ostatecznie, liczba ulegających ekspresji genów MHC (na podstawie czterech replikatów) została ustalona dla 72 osobników. Z pośród nich, wybrałam do finalnego etapu – sekwencjonowania repertuaru limfocytów TCRβ – 30 osobników. Kryteriami wybory była wysoka jakość wyekstrahowanego RNA, równa reprezentacja płci, osobników o wysokiej i niskiej liczbie wariantów MHC, a także kierunków selekcji z bazalnej kolonii. Biblioteki do sekwencjonowania repertuaru TCR zostały przygotowane tak, jak przy opisie ilościowym prezentowanym w drugiej części

niniejszej rozprawy. W wyniku sekwencjonowania uzyskano odpowiednią liczbę odczytów (1 mln) dla 28 analizowanych osobników, dla których szacowana była wielkość repertuaru CDR3 TCRβ z wykorzystaniem estymatora Chao2. Estymowana wielkość repertuaru była predyktorem w mieszanym modelu liniowym, w którym zmiennymi objaśniającymi (efekty stałe) były: liczba supertypów/wariantów MHC klasy I, klasy II, płeć i kierunek selekcji. Dodatkowo brane był pod uwagę następujące efekty losowe: data śmierci zwierzęcia (i poboru organów) oraz linia wewnątrz kierunku selekcji. Analizowanie obu klas MHC jednocześnie pozwoliło na wykazanie nieoczekiwanych różnic pomiędzy nimi – które nigdy wcześniej nie były uwzględniane w modelach matematycznych hipotezy optymalności, ani w pośrednich testach tej hipotezy. Analiza statystyczna wykazała negatywny wpływ wysokiej liczby supertypów/wariantów aminokwasowych MHC klasy I na wielkość repertuaru TCR, oraz brak wpływu liczby wariantów MHC klasy II. Ponadto, na wielkość repertuaru TCR wpływ miała płeć – samce charakteryzowały się istotnie mniejszym repertuarem TCR niż samice. Wynik ten częściowo wspiera hipotezę optymalności, jednocześnie stawiając nowe pytania badawcze – obecnie nie wiadomo, co powoduje obserwowaną różnicę we wpływie poszczególnych klas MHC na repertuar TCR. Niniejsza praca jest więc ważnym krokiem w kierunku zrozumienia kompromisów ewolucyjnych kształtujących układ odpornościowy kręgowców. Podważa ona uniwersalność dotąd powszechnie przyjmowanej hipotezy, zwracając uwagę na konieczność uwzględnienia w jej ramach pomijanych lub nadmiernie uogólnianych aspektów immunologicznej złożoności. Praca ta również podkreśla wagę i wartość empirycznych testów klasycznych, ewolucyjnych teorii.

II. Summary

The classical major histocompatibility complex (MHC) genes encode molecules that present antigens to cells mediating acquired immune response of vertebrates: lymphocytes T, also called T cells (Klein 1986). There are two types of classical MHC molecules, class I and class II, that differ in structure, tissue distribution and function. MHC class I molecules are composed of a heavy α chain and a light mβ2 molecule. They are found on all nucleated cells, and generally present cytosolic peptides (e.g., from viruses or intracellular bacteria) to cytotoxic T cells. MHC class II molecules are composed of two similar chains, α and β, and are found on specialized, antigen presenting cells (e.g. dendritic cells, B cells, macrophages). They generally present peptides derived from pathogens found in the extracellular space to helper T cells (Klein 1986; Kindt et al. 2007).

While MHC is necessary to bind and present antigens, a primary response of the immune system comes from αβ T cells. The response is initiated upon a specific recognition of an MHC-bound antigen, by a T cell receptor (TCR). TCRs are heterodimeric receptors, build of α and β chain. Each chain is composed of a Constant and a Variable region. The Variable region is assembled, similarly to antibodies, through the process of somatic recombination of gene segments: V (*variable*) and J (*joining*) in the α chain; V, D (*diversity*) and J in the β chain. Apart from the combinatorial diversity resulting from an association of different V, D and J segments present in the genome, additional variation arises form addition and deletion of random nucleotides at the junctions between these segments – so-called N-diversity regions. The resulting, hypervariable part of the chain, formed at the V(D)J junctions, encodes third complementarity-determining region, CDR3, that specifically recognizes peptide/MHC complex (Davis and Bjorkman 1988; Kindt et al. 2007).

The interaction between the peptide/MHC complexes and TCRs forms the basis for molecular self/non-self recognition. In a healthy organism, detection of a foreign antigen leads to the initiation of an adaptive immune response (Kindt et al. 2007). Self-peptides, however, are ignored, as a result of central tolerance acquired during T cell maturation in thymus. The two main processes ensuring central tolerance are positive and negative selection (Klein et al. 2009, 2014). The positive selection promotes

survival of T cells that are able to interact with MHC/self-peptide complexes, while lack of such interaction leads to death of neglect. During negative section, T cells bearing TCRs binding MHC/self-peptides with a too strong avidity are deleted, to prevent autoimmune responses (Klein et al. 2014). Due to a large, stochastic component during formation of TCRs, vast majority of the initially generated TCR diversity is either non-responsive or strongly self-reactive, and is further removed during T cell maturation in thymus. It is estimated that 20-25% of the cell beginning thymic education is positively selected, of which 20-50% survives negative selection. Overall, as little as 5% of T cells completes the thymic education (reviewed in Yates 2014). These pruning processes assure a diverse and functional TCR repertoire of mature T cells, that are MHC-restricted and self-tolerant.

My PhD thesis focus on a hypothesis, that proposes the negative thymic selection as a factor limiting the evolutionary, within-individual expansion and diversification of the classical MHC genes. The immense polymorphism found at MHC was a subject of evolutionary research for decades. Usually, dozens or hundreds of alleles are found in vertebrate populations, and in extreme cases (e.g., human population) thousands of alleles have been described (Apanius et al. 1997; Bernatchez and Landry 2003). It is commonly accepted now, that this extraordinary diversity results from an evolutionary "arms race" between hosts and pathogens (Jeffery and Bangham 2000; Spurgin and Richardson 2010). Differences among alleles accumulate in the peptide-binding clefts of the MHC molecules, with characteristically increased rate of nonsynonymous to synonymous substitutions (Reche and Reinherz 2003). This pattern is interpreted as positive selection acting to diversify amino acid sequences of the molecules, so that different alleles could bind different peptides. Further, the great allelic diversity, is maintained in populations by various forms of balancing selection. This includes, but is not limited to, heterozygote advantage (increased fitness of heterozygotes, brought about by a wider spectrum of recognized pathogens; Doherty and Zinkernagel 1975; Penn et al. 2002; Oliver et al. 2009), and negative frequency-dependent selection (increased fitness of individuals bearing rare or novel alleles, caused by an adaptation of pathogens to evade recognition by prevailing MHC variants; Borghans et al. 2004; Ejsmond and Radwan 2015; Phillips et al. 2018).

While a lot of effort was devoted to the explanation of mechanisms behind generation and maintenance of the MHC polymorphism at the population level

(Bernatchez and Landry 2003; Spurgin and Richardson 2010), much less attention was paid to the mechanisms shaping the MHC gene number at the individual level. Given the adaptive value of the MHC polymorphisms, an advantage of heterozygosity and an evolutionary potential for gene duplication – rather limited number of MHC loci within individual genomes is surprising. While there is usually a few functional loci, they accommodate only a small fraction of the (presumably adaptive) MHC diversity present at the population level. An explanation for this apparent paradox was proposed over 30 years ago (Vidović and Matzinger 1988), and later formalized as a mathematical model of immunogenetic optimality, or an *optimality hypothesis* (Nowak et al. 1992). It predicts an evolutionary trade-off between the ability to present a wider spectrum of antigens (brought about by an expansion of MHC genes within genome), and a decreased T cell diversity (caused by mechanisms conferring central tolerance). This is because a broadened spectrum of bound peptides would also include self-peptides, and a larger fraction of TCRs might turn out self-reactive and be pruned in the process of negative selection; hence another term: *TCR depletion hypothesis* (Woelfing et al. 2009). However, a critique of this hypothesis was put forward, building on a notion that the broadened spectrum of bound self-peptides should, in the first place, increase cell survival at the positive selection step (Borghans et al. 2003). Therefore, depending on the parameters used, theory predicts that an increase in the individual number of MHC genes will either lead to a decrease (Nowak et al. 1992; Woelfing et al. 2009), or a relative increase in the TCR repertoire size (Borghans et al. 2003). Resolution of the debate have been hampered by a persistent uncertainty surrounding several key parameters and assumptions underlying the abovementioned mathematical models (Yates 2014; La Gruta et al. 2018). In particular, it is yet not fully understood, to a what extend the positive selection is driven by an interaction between less variable and germline-encoded parts of the MHC and TCRs, versus by an interaction between diverse peptides alongside the highly polymorphic peptide-binding cleft of the MHC and the hypervariable CDR3s of TCRs (Scott-Browne et al. 2009; Holland et al. 2012; Baker and Evavold 2017).

Ecologist and evolutionary biologist tackled the issue of an immunogenetic optimality with a different approach, and worked towards providing indirect evidence in support of the optimality hypothesis. In their research, they used species characterized by an intra-specific variation in the number of MHC genes, – a phenomena commonly referred to as copy number variation (CNV), though it does not involve actual, identical

copies, but rather a variable number of duplicated and diversified loci (Siddle et al. 2010; Lighten et al. 2014). They assumed that individuals with an average – optimal – number of MHC genes, would have the highest immunocompetence or fitness. However, studies correlating the individual MHC gene number with e.g., parasite diversity or parasite load, lifelong reproductive success or other fitness correlates yielded mixed results (Wegner et al. 2003; Bonneaud et al. 2004; Madsen and Ujvardi 2006; Radwan et al. 2012; Westerdahl et al. 2012; Sepil et al. 2013; Biedrzycka et al. 2018). The indirect nature of these tests might be responsible for the inconclusiveness of the results. There are numerous factors affecting infection outcome or fitness, such as an effect of particular MHC alleles or differences in spatiotemporal parasite pressure (O'Connor et al. 2018).

For a long time, a direct test of the predictions of the optimality hypothesis was missing, despite the clear importance of this hypothesis for our understanding of the evolution of the adaptive immune system of vertebrates, and its implications for the evolutionary processes well beyond immunology, such as speciation (Eizaguirre et al. 2009; Malmstrøm et al. 2016), In particular, a negative relationship between a high individual MHC diversity/gene number, and the TCR repertoire size, has never been shown. The lack of evidence resulted mainly from technical difficulties related to sequencing and analysis of the immense immune repertoires, but this obstacle was finally removed with an advent of the next generation, high throughput sequencing (HTS) (Benichou et al. 2012).

The main aim of this PhD thesis was to **test a key protection of *the optimality hypothesis*: whether individuals with a high number of MHC genes indeed have a smaller size of the TCR repertoire**. The study species was the bank vole (*Myodes glareolus*), a small rodent from the Cricetidae family, often used in evolutionary and ecological studies. Bank voles used in this project were obtained from a laboratory colony held at the Institute of the Environmental Sciences, Jagiellonian University, Kraków (courtesy of prof. P. Koteja). This species was chosen mainly because of an extensive, between individual variation in the number of expressed MHC genes (previously well-documented in the MHC class II, e.g., Axtner & Sommer, 2007; Bryja et al., 2006). However, as it is not a standard model in molecular biology, it lacked several basic, immunogenetic resources. Therefore, in order to accomplish the main research goal, **it was necessary to characterize crucial genes involved in**

antigen presentation and recognition. In particular, a description of MHC diversity (primarily of MHC class I, but also of the expressed CNV of MHC class II), and an estimation of the TCR repertoire size and diversity was missing.

In the first part of my PhD thesis, I characterized expressed MHC class I genes in the bank vole (Migalska et al. 2017). Before, only orthologues of MHC class II were characterized in this species: DQA (Bryja et al. 2006; Deter et al. 2008), DQB (Scherman et al. 2014) and DRB (Axtner and Sommer 2007; Kloch et al. 2010). I used *de novo* assembled spleen transcriptomes of seven bank voles to design specific primers. Next, I amplified coding sequence (CDS) of MHC class I (exons 1-8 and part of 3'UTR), cloned in into bacterial vectors and sequenced using Sanger sequencing. I used exons 2-4 (coding for the extracellular part of the MHC molecule), to conduct a phylogenetic analysis of a relationship between bank vole alleles and those of other species from Cricetidae and Muridae family. The analysis revealed lack of orthology, typical for a higher gene turnover of MHC class I genes, compared to MHC class II. Moreover, to obtain a more detailed data on polymorphism, I sequenced with HTS exon three of MHC class I, in 29 individuals. I used AmpliSAS (Sebastian et al. 2016) – a program developed in the Evolutionary Biology Group (EBP) from AMU University – for a multi-locus MHC genotyping. Even in this limited sample, considerable variation in the number of expressed MHC genes was present, ranging from 5 and 14 distinct variants. Finally, I analyzed signatures of selection, by comparison of the rate of synonymous to nonsynonymous substitutions. It revealed eight codons under positive selection, most of which overlap with antigen-binding residues in human.

In the second part of my thesis, I described V and J segments building TCRβ chains, characterized CDR3 length distribution and estimated the TCRβ repertoire size in the bank vole (Migalska et al. 2018). A qualitative part of the study (based on seven individuals) described Variable part of TCRβ chains, which encompass recombined V, D and J segments. I applied 5' RACE (*Rapid Amplification of cDNA Ends*) technique, that uses one primer complementary to the 5' end of the Constant region, and a universal adapter incorporated at the 5' end of the TCRβ cDNA (Mamedov et al. 2013). Nested primers for 5′ RACE were designed based on *de novo* assembled transcriptomes, as described previously for MHC class I. The obtained template was amplified with universal primers and sequenced with HTS on Illumina MiSeq platform (read length 2×300 bp, sufficient to cover the entire sequence). I used AmpliTCR and AmpliCDR3, programs developed in the EBG AMU (Migalska et al. 2018), for

bioinformatics analysis. I identified 37 groups of V and 11 groups of J genes; and these groups likely represent loci. Phylogenetic analysis revealed preservation of an overall orthology to the murine genes, despite considerable diversification. Further, I described CDR3 length distribution and preferential V–J segment usage – for the first time in the bank vole, and in general, it was a first analysis of this kind in a non-model species. The quantitative part of this study (based on three individuals), estimated a lower boundary of the TCRβ repertoire size in the bank vole. To this aim, I used a modified 5' RACE technique, with 5' adapters incorporating unique molecular identifiers (UMIs) (Shugay et al. 2014). The use of UMIs allows effective discrimination and elimination of PCR and sequencing errors, particularly important in analyses of the immense immune repertoires. I made four PCR replicates for each individual, and sequenced them at high depth (>1 mln reads) on Illumina HiSeq2500. The bioinformatics was done with AmpliCDR3. Further, I applied a richness estimator derived from ecological census techniques - Chao 2 (Chao 1987), - to estimate a variation in the TCR repertoire size in this species. I used a subsample of 1.5 million sequencing reads, to avoid bias introduced by unequal sequencing depths. Mean number of unique TCRβ CDR3 sequenced per individual (combined from four replicates) was 1.8×10^5 of amino acid sequences, and 1.5×10^5 of amino acid sequences. Lower bound of TCR repertoire size was estimated at 2.1×10^5 nucleotide sequences.

Third part of my thesis presents the results of a direct test of predictions of the optimality hypothesis (Migalska et al. 2019)**.** First, I determined the variability in the number of expressed MHC class I and class II variants in ca. 150 individuals. I prepared two independent PCR replicates for each marker (i.e., gene fragment amplified with a pair of primers), sequenced them with PGM Ion Torrent and genotyped with a modified genotyping method from the first part of the thesis. This preliminary results confirmed high between-individual variation in number of MHC class II genes in this species, and showed that it holds at the RNA level (DQA: 2-6, DQB: 2-8, DRB: 1-8). Next, I chose approximately 40 individuals with low and 40 individuals with high number of MHC variants, and for them, I prepared two additional PCR replicates for each marker (sequenced with Illumina MiSeq), to increase precision of MHC typing. I successfully estimated number of expressed MHC genes (based on four PCR replicates) for 72 individuals (out of 77 sequenced). From these, I chose 30 individuals for the final, experimental stage - TCRβ repertoire sequencing. The individuals were chosen based on a high quality of extracted RNA, and to equally represent sexes, selection

directions from the basal colony and individuals from low and high MHC-variant number groups. Libraries for HTS were prepared as in the second section of this thesis (quantitative part). 28 individuals reached a desired sequencing depth (1 million reads). Their estimated TCRβ CDR3 repertoire size was used as a predictor in a mixed-effect linear model, with the following explanatory variables (fixed effects): number of MHC class I supertypes/variants, number of MHC class II supertypes/variants, sex, selection direction. The model contained also the following random effects: date of death/organ collection and line within selection direction. Use of the MHC classes as separate predictors exposed a discrepancy between them - unanticipated by the models, and missed by the indirect tests. It showed a negative correlation between the TCR repertoire size and the number of MHC class I diversity, but no correlation for MHC class II. Another factor shown to affect the TCR repertoire size was sex: males had a significantly smaller TCR repertoire than females. In summary, this work was the first, direct test of an important, evolutionary hypothesis for the mechanism limiting within-individual MHC diversity. The result partially supported the optimality hypothesis, but at the same time, raised new research questions – we still do not know what causes the observed disparity between the two MHC classes in their effect on TCR diversity. This work is therefore an important step towards understanding the evolutionary trade-offs shaping the immune system of vertebrates. It challenges the generality of a commonly accepted hypothesis, highlighting the need to account for a previously neglected aspects of immunological complexity. This work also emphasizes the importance and value of empirical tests of classical, evolutionary theories.

References

Apanius V, Penn D, Slev PR, et al (1997) The nature of selection on the major histocompatibility complex. Crit Rev Immunol 17:179–224

Axtner J, Sommer S (2007) Gene duplication, allelic diversity, selection processes and adaptive value of MHC class II DRB genes of the bank vole, Clethrionomys glareolus. Immunogenetics 59:417–426. doi: 10.1007/s00251-007-0205-y

Baker BM, Evavold BD (2017) MHC Bias by T Cell Receptors: Genetic Evidence for MHC and TCR Coevolution. Trends Immunol 38:2–4. doi: 10.1016/j.it.2016.11.003

Benichou J, Ben-Hamo R, Louzoun Y, Efroni S (2012) Rep-Seq: Uncovering the immunological repertoire through next-generation sequencing. Immunology 135:183–191. doi: 10.1111/j.1365-2567.2011.03527.x

Bernatchez L, Landry C (2003) MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? J Evol Biol 16:363–377. doi: 10.1046/j.1420-9101.2003.00531.x

Biedrzycka A, Bielański W, Ćmiel A, et al (2018) Blood parasites shape extreme major histocompatibility complex diversity in a migratory passerine. Mol Ecol 27:2594–2603. doi: 10.1111/mec.14592

Bonneaud C, Mazuc J, Chastel O, et al (2004) Terminal Investment Induced by Immune Challenge and Fitness Traits Associated with Major Histocompatibility Complex in the House Sparrow. Evolution (N Y) 58:2823–2830. doi: 10.1111/j.0014-3820.2004.tb01633.x

Borghans JAM, Beltman JB, De Boer RJ (2004) MHC polymorphism under host-pathogen coevolution. Immunogenetics 55:732–9. doi: 10.1007/s00251-003-0630-5

Borghans JAM, Noest AJ, De Boer RJ (2003) Thymic selection does not limit the individual MHC diversity. Eur J Immunol 33:3353–8. doi: 10.1002/eji.200324365

Bryja J, Galan M, Charbonnel N, Cosson JF (2006) Duplication, balancing selection and trans-species evolution explain the high levels of polymorphism of the DQA MHC class II gene in voles (Arvicolinae). Immunogenetics 58:191–202. doi: 10.1007/s00251-006-0085-6

Chao A (1987) Estimating the population size for capture-recapture data with unequal catchability. Biometrics 43:783–791

Davis MM, Bjorkman PJ (1988) T-cell antigen receptor genes and T-cell recognition. Nature 334:395–402. doi: 10.1038/334395a0

Deter J, Bryja J, Chaval Y, et al (2008) Association between the DQA MHC class II gene and Puumala virus infection in Myodes glareolus, the bank vole. Infect Genet Evol 8:450–8. doi: 10.1016/j.meegid.2007.07.003

Doherty PC, Zinkernagel RM (1975) Enhanced immunological surveillance in mice

heterozygous at the H-2 gene complex. Nature 256:50–52. doi: 10.1038/256050a0

Eizaguirre C, Lenz TL, Traulsen A, Milinski M (2009) Speciation accelerated and stabilized by pleiotropic major histocompatibility complex immunogenes. Ecol Lett 12:5–12. doi: 10.1111/j.1461-0248.2008.01247.x

Ejsmond MJ, Radwan J (2015) Red Queen Processes Drive Positive Selection on Major Histocompatibility Complex (MHC) Genes. PLoS Comput Biol 11:e1004627. doi: 10.1371/journal.pcbi.1004627

Holland SJ, Bartok I, Attaf M, et al (2012) The T-cell receptor is not hardwired to engage MHC ligands. Proc Natl Acad Sci U S A 109:E3111-8. doi: 10.1073/pnas.1210882109

Jeffery KJM, Bangham CRM (2000) Do infectious diseases drive MHC diversity? Microbes Infect 2:1335–1341. doi: 10.1016/S1286-4579(00)01287-9

Kindt TJ, Goldsby RA, Osborne BA, Kuby J (2007) Kuby Immunology. W. H. Freeman

Klein J (1986) Natural history of the major histocompatibility complex. Wiley, New York

Klein L, Hinterberger M, Wirnsberger G, Kyewski B (2009) Antigen presentation in the thymus for positive selection and central tolerance induction. Nat Rev Immunol 9:833–44. doi: 10.1038/nri2669

Klein L, Kyewski B, Allen PM, Hogquist KA (2014) Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). Nat Rev Immunol 14:377–91. doi: 10.1038/nri3667

Kloch A, Babik W, Bajer A, et al (2010) Effects of an MHC-DRB genotype and allele number on the load of gut parasites in the bank vole Myodes glareolus. Mol Ecol 19 Suppl 1:255–65. doi: 10.1111/j.1365-294X.2009.04476.x

La Gruta NL, Gras S, Daley SR, et al (2018) Understanding the drivers of MHC restriction of T cell receptors. Nat Rev Immunol 18:467–478. doi: 10.1038/s41577-018-0007-5

Lighten J, van Oosterhout C, Paterson IG, et al (2014) Ultra-deep Illumina sequencing accurately identifies MHC class IIb alleles and provides evidence for copy number variation in the guppy (Poecilia reticulata). Mol Ecol Resour 14:753–67. doi: 10.1111/1755-0998.12225

Madsen T, Ujvardi B (2006) MHC class I variation associates with parasite resistance and longevity in tropical pythons. J Evol Biol 19:1973–1978. doi: 10.1111/j.1420-9101.2006.01158.x

Malmstrøm M, Matschiner M, Tørresen OK, et al (2016) Evolution of the immune system influences speciation rates in teleost fishes. Nat Genet 48:1204–1210. doi: 10.1038/ng.3645

Mamedov IZ, Britanova O V, Zvyagin I V, et al (2013) Preparing unbiased T-cell receptor and antibody cDNA libraries for the deep next generation sequencing profiling. Front Immunol 4:456. doi: 10.3389/fimmu.2013.00456

Migalska M, Sebastian A, Konczal M, et al (2017) De novo transcriptome assembly facilitates characterisation of fast-evolving gene families, MHC class I in the bank vole (Myodes glareolus). Heredity (Edinb) 118:348–357. doi: 10.1038/hdy.2016.105

Migalska M, Sebastian A, Radwan J (2018) Profiling of the TCRβ repertoire in non-model species using high-throughput sequencing. Sci Rep 8:11613. doi: 10.1038/s41598-018-30037-0

Migalska M, Sebastian A, Radwan J (2019) Major histocompatibility complex class I diversity limits the repertoire of T cell receptors. Proc Natl Acad Sci 201807864. doi: 10.1073/PNAS.1807864116

Nowak MA, Tarczy-Hornoch K, Austyn JM (1992) The optimal number of major histocompatibility complex molecules in an individual. Proc Natl Acad Sci 89:10896–10899. doi: 10.1073/pnas.89.22.10896

O'Connor EA, Cornwallis CK, Hasselquist D, et al (2018) The evolution of immunity in relation to colonization and migration. Nat Ecol Evol 2:841–849. doi: 10.1038/s41559-018-0509-3

Oliver M., Telfer S, Piertney S. (2009) Major histocompatibility complex (MHC) heterozygote superiority to natural multi-parasite infections in the water vole (Arvicola terrestris). Proc R Soc B Biol Sci 276:1119–1128. doi: 10.1098/rspb.2008.1525

Penn DJ, Damjanovich K, Potts WK (2002) MHC heterozygosity confers a selective advantage against multiple-strain infections. Proc Natl Acad Sci U S A 99:11260–11264. doi: 10.1073/pnas.162006499

Phillips KP, Cable J, Mohammed RS, et al (2018) Immunogenetic novelty confers a selective advantage in host-pathogen coevolution. Proc Natl Acad Sci U S A 201708597. doi: 10.1073/pnas.1708597115

Radwan J, Zagalska-Neubauer M, Cichoń M, et al (2012) MHC diversity, malaria and lifetime reproductive success in collared flycatchers. Mol Ecol 21:2469–2479. doi: 10.1111/j.1365-294X.2012.05547.x

Reche P a., Reinherz EL (2003) Sequence variability analysis of human class I and class II MHC molecules: Functional and structural correlates of amino acid polymorphisms. J Mol Biol 331:623–641. doi: 10.1016/S0022-2836(03)00750-2

Scherman K, Råberg L, Westerdahl H (2014) Positive Selection on MHC Class II DRB and DQB Genes in the Bank Vole (Myodes glareolus). J Mol Evol 78:293–305. doi: 10.1007/s00239-014-9618-z

Scott-Browne JP, White J, Kappler JW, et al (2009) Germline-encoded amino acids in the αβ T-cell receptor control thymic selection. Nature 458:1043–1046. doi: 10.1038/nature07812

Sebastian A, Herdegen M, Migalska M, Radwan J (2016) amplisas: a web server for multilocus genotyping using next-generation amplicon sequencing data. Mol Ecol Resour 16:498–

510. doi: 10.1111/1755-0998.12453

Sepil I, Lachish S, Hinks AE, Sheldon BC (2013) Mhc supertypes confer both qualitative and quantitative resistance to avian malaria infections in a wild bird population. Proc R Soc B Biol Sci 280:20130134–20130134. doi: 10.1098/rspb.2013.0134

Shugay M, Britanova O V, Merzlyak EM, et al (2014) Towards error-free profiling of immune repertoires. Nat Methods 11:653–5. doi: 10.1038/nmeth.2960

Siddle H V., Marzec J, Cheng Y, et al (2010) MHC gene copy number variation in Tasmanian devils: implications for the spread of a contagious cancer. Proc R Soc B Biol Sci 277:2001–2006. doi: 10.1098/rspb.2009.2362

Spurgin LG, Richardson DS (2010) How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proc Biol Sci 277:979–88. doi: 10.1098/rspb.2009.2084

Vidović D, Matzinger P (1988) Unresponsiveness to a foreign antigen can be caused by self-tolerance. Nature 336:222–225. doi: 10.1038/336222a0

Wegner KM, Kalbe M, Kurtz J, et al (2003) Parasite selection for immunogenetic optimality. Science 301:1343. doi: 10.1126/science.1088293

Westerdahl H, Asghar M, Hasselquist D, Bensch S (2012) Quantitative disease resistance: to better understand parasite-mediated selection on major histocompatibility complex. Proc Biol Sci 279:577–84. doi: 10.1098/rspb.2011.0917

Woelfing B, Traulsen A, Milinski M, Boehm T (2009) Does intra-individual major histocompatibility complex diversity keep a golden mean? Philos Trans R Soc Lond B Biol Sci 364:117–28. doi: 10.1098/rstb.2008.0174

Yates AJ (2014) Theories and quantification of thymic selection. Front Immunol 5:13. doi: 10.3389/fimmu.2014.00013