Gut Virome Dynamics During Pregnancy in Mothers with Type 1 Diabetes in the ENDIA Study

*Korresponding Authors: Dr. Kinya Wee Kim (k-w.kim@unsw.edu.au) and Prof. Maria Craig (m.craig@unsw.edu.au)
Prince of Wales Hospital & UNSW Virology Research Laboratory, University of New South Wales, Sydney, NSW, Australia

Kinya Wee Kim*, Digby W. Allen, Thomas Briese, Chi Nam Pang, Komal Jain, Jessica L. Horton, Sonia R. Isaacs, Marc R. Wilkins, Jenny J. Couper, Megan A.S. Penno, Leonard C. Harrison, Peter Colman, Rebecca L. Thomson, Grant Morahan, John Wentworth, Mark Harris, Elizabeth Davis, Aven Haynes, Lynne Giles, Richard Sinnott, Peter Vuillemir, Andrew M. Cotterill, Claire Morby, Georgia Soldatos, Jodie Dodd, Simon C. Barry, Jackie Catteau, Kelly Spooner, Debra Beuzidenhout, Bernadette Rice, Emily Gibson, Jan Lipkin, William D. Rawlinson, Maria E. Craig*, ENDIA Study Group

Introduction
- Rapid rise in global incidence, discordance in monozygotic twins, seasonal/geographical variations all support environmental trigger(s).
- Virus infection is a prime candidate for the environmental trigger of T1D4
- ENDIA is an Australian prospective T1D birth cohort study, following 1400 genetically predisposed children from pregnancy to first three years of life2
- ENDIA is multifaceted, investigating associations of a diverse spectrum of environmental factors, including the "virome" (population of all viruses)
- Maternal virus infections during pregnancy may increase the risk of T1D in the offspring4
- However, very little is known about the diversity and dynamics of the virome in healthy women during pregnancy, let alone in mothers with T1D.

To characterise the virome of ENDIA participant mothers to answer:
Q1. What is the prevalence of maternal virus infection during pregnancy?
Q2. Does the gut virome differ between mothers with and without T1D?

Methods
- Virome capture sequencing of vutebrate-infecting viruses (ViCapSeq-VERT) was performed on 124 longitudinal stool specimens collected across three different trimester periods from 34 mothers with T1D (n=68) and 27 without T1D (n=56).
- Workflow summarised below (Fig. 1).
- Trimester definitions (weeks): T1 = 1-14; T2 = 15-26; T3 = 27-42

Table 1. Summary of maternal characteristics. Data are expressed as means ± SD or n (%) of patients. P-values calculated using two-tailed t-test.

<table>
<thead>
<tr>
<th>Maternal characteristics</th>
<th>Without T1D (n=26)</th>
<th>With T1D (n=35)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DR24</td>
<td>0</td>
<td>6 (22.2)</td>
<td>1</td>
</tr>
<tr>
<td>HLA status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DR3X</td>
<td>0</td>
<td>9 (33.3)</td>
<td>0.5</td>
</tr>
<tr>
<td>DR4X</td>
<td>2 (66.6)</td>
<td>10 (37)</td>
<td>0.5</td>
</tr>
<tr>
<td>DRRX</td>
<td>1 (33.3)</td>
<td>2 (7.4)</td>
<td>0.3</td>
</tr>
<tr>
<td>Age at conception (years)</td>
<td>33.8 ± 4.1</td>
<td>31 ± 4.5</td>
<td>0.02</td>
</tr>
<tr>
<td>Socioeconomic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Status</td>
<td>Low (1st quartile)</td>
<td>0 (0)</td>
<td>0.3</td>
</tr>
<tr>
<td>Modum (2nd quartile)</td>
<td>10 (38.5)</td>
<td>15 (42.9)</td>
<td>0.8</td>
</tr>
<tr>
<td>High (3rd quartile)</td>
<td>16 (61.5)</td>
<td>17 (48.6)</td>
<td>0.4</td>
</tr>
<tr>
<td>Pre-pregnancy BMI</td>
<td>26.5 ± 6.1</td>
<td>27.1 ± 5.9</td>
<td>0.95</td>
</tr>
<tr>
<td>Smoking History</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never Smoked</td>
<td>19 (76)</td>
<td>26 (74.3)</td>
<td>1</td>
</tr>
<tr>
<td>Smoking History</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Previous Smoker</td>
<td>5 (20)</td>
<td>9 (25.7)</td>
<td>0.8</td>
</tr>
<tr>
<td>Current Smoker</td>
<td>1 (4)</td>
<td>0</td>
<td>0.4</td>
</tr>
<tr>
<td>Pots</td>
<td>19 (76)</td>
<td>23 (67.6)</td>
<td>0.6</td>
</tr>
<tr>
<td>Comorbidities</td>
<td>2.4 (1.7)</td>
<td>1.9 ± 1.7</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Figure 1. Workflow of ENDIA virome capture sequencing and bioinformatics pipeline: Key kits, software and tools are listed. Work performed in collaboration with Center for Infection and Immunology at Columbia University, NY.

Figure 2. Frequency of viruses detected by ViCapSeq-VERT in 124 pregnancy stool specimens based on 100 viral read positivity threshold (p<0.006).

Figure 3. Heatmap of viral reads detected by ViCapSeq-VERT in mothers with and without T1D. Only mothers with multiple longitudinal specimens are included to illustrate dynamics of the gut virome over time. Viruses detected in three or more specimens are represented at the genus level. Positivity threshold of 100 viral reads matched by BLAST at the species level was applied.

Figure 4. Examples of high virus genome coverage achieved using ViCapSeq-VERT. Filtered reads were mapped to reference genome accessions identified by BLAST (megablast). Partial/complete genome sequences were obtained from GenBank, alignment performed using bowtie2 and coverage plots generated in R package ViralG.

Figure 5. Volcano plot of differentially abundant virus species in the gut between mothers with and without T1D. Only viruses with ≥2-fold difference (dotted line) and FDR below 5% (q<0.05) based on edgeR are represented.

Conclusions
- This is the first study to utilise virome capture sequencing in any pregnancy or case-control study
- Population and abundance of viruses in the gut differ significantly between women with and without T1D
- Whether these differences influence the risk of IA/T1D in the offspring will be examined by longitudinal follow up
- This work has been published and available via Open Access in the Open Forum Infectious Diseases Journal.

References
The virome in early life and childhood and development of islet autoimmunity and type 1 diabetes: A systematic review and meta-analysis of observational studies

Clare L. Faulkner1,2 | Yi Xuan Luo1,2 | Sonia Isaacs1,2 | William D. Rawlinson1,2,3,4 | Maria E. Craig1,2,5,6 | Ki Wook Kim1,2

Summary
Viruses are postulated as primary candidate triggers of islet autoimmunity (IA) and type 1 diabetes (T1D), based on considerable epidemiological and experimental evidence. Recent studies have investigated the association between all viruses (the ‘virome’) and IA/T1D using metagenomic next-generation sequencing (mNGS). Current associations between the early life virome and the development of IA/T1D were analysed in a systematic review and meta-analysis of human observational studies from Medline and EMBASE (published 2000–June 2020), without language restriction. Inclusion criteria were as follows: cohort and case–control studies examining the virome using mNGS in clinical specimens of children ≤18 years who developed IA/T1D. The National Health and Medical Research Council level of evidence scale and Newcastle–Ottawa scale were used for study appraisal. Meta-analysis for exposure to specific viruses was performed using random-effects models, and the strength of association was measured using odds ratios (ORs) and 95% confidence intervals (CIs). Eligible studies (one case–control, nine nested case–control) included 1,425 participants (695 cases, 730 controls) and examined IA (n = 1,023) or T1D (n = 402). Meta-analysis identified small but significant associations between IA and number of stool samples positive for all enteroviruses (OR 1.14, 95% CI 1.00–1.29, p = 0.05; heterogeneity χ² = 1.51, p = 0.03; I² = 0%), consecutive positivity for enteroviruses (1.55, 1.09–2.20, p = 0.01; χ² = 0.19, p = 0.91, I² = 0%) and number of stool samples positive specifically for enterovirus B (1.20, 1.01–1.42, p = 0.04; χ² = 0.03, p = 0.86, I² = 0%).
Type 1 diabetes (T1D) is common, affecting more than 600,000 children aged <15 years worldwide.\(^1\) T1D is preceded by islet autoimmunity (IA) lasting months to decades.\(^2\) It is defined serologically as multiple autoantibodies against one or more T1D-associated autoantigens, including insulin (IAA), glucatmic-acid decarboxylase (GADA), tyrosine phosphatase-like insulinoma antigen 2 (IA2A), islet cell cytoplasmic proteins (ICA) and \(\beta\)-cell-specific zinc transporter 8 (ZnT8A).\(^3\) T1D pathogenesis results from a complex interplay of genetic predisposition\(^4,5\) and environmental exposures.\(^6,7\) Accumulating evidence supports the influence of environmental factors, particularly viruses. The increased incidence of T1D is too high to be attributed to genetics alone,\(^8,9\) with data showing seasonal IA/T1D clustering,\(^9\) geographical variation in incidence\(^10\) and more frequent in utero and early-life infections in affected individuals.\(^11,12\)

Higher rates of enterovirus (EV) infection, detected by serological\(^13,14\) or molecular methods,\(^15–17\) have been observed in T1D patients at diagnosis versus unaffected controls, or prospectively in individuals who subsequently develop IA and/or T1D versus those who do not. Accordingly, our previous meta-analysis investigating EV using molecular methods demonstrated significant association between EV and IA (odds ratio [OR] 3.7, 95% confidence interval [CI] 2.1–6.8, \(p < 0.001\)) or T1D (9.8, 5.5–17.4, \(p < 0.001\)).\(^18\) In addition, EV proteins and RNA have been isolated from pancreata of affected patients, with upregulated EV receptors selectively expressed in pancreatic islets.\(^19,20\) However, inconsistencies in findings\(^21–23\) make it difficult to establish a definitive causal association. Importantly, substantial investigation bias exists for EVs in previous studies.\(^22\) In contrast, only a limited number of studies have reported on the potential associations of other viruses with T1D, including mumps,\(^24\) cytomegalovirus,\(^25\) rotavirus,\(^26\) parechovirus,\(^27,28\) Epstein–Barr virus,\(^29\) rubella\(^12\) and parvovirus.\(^30\)

In an effort to alleviate this bias towards EVs, a growing number of studies are applying high-throughput metagenomic next-generation sequencing (mNGS) to comprehensively characterise the population of all known human viruses (the ‘virome’), simultaneously. Here, we report the first systematic review and meta-analysis of observational studies using mNGS to investigate vertebrate-infecting DNA and RNA viruses in children \(\leq 18\) years, and subsequent development of IA or T1D. Analysis of bacteriophage has been excluded from this review. The unbiased viral mNGS in early life and childhood has potential to comprehensively identify diabetogenic viruses increasing the IA/T1D risk or viruses affording protection. This may present new opportunities to intervene through antiviral medications or vaccination.

2 | METHODS

2.1 | Search strategy and selection criteria

This review is registered on PROSPERO (23 July 2020), registration number CRD42020188737. Two reviewers (Clare L. Faulkner and Yi Xuan Luo) independently conducted a systematic search for observational studies investigating the association between virome composition and/or abundance, and IA or T1D. EMBASE and MEDLINE databases were searched (2000–1 June 2020) using the strategy in Appendix S1. The search was performed without geographical or language restrictions and limited to studies in humans. Restriction to studies published from year 2000 onwards was informed by emergence of mNGS and other viral sequencing technologies.\(^31,32\) This search was supplemented by manual searching of references of identified papers, key journals, OpenGrey and ProQuest to identify additional articles potentially missed by online indexes. PROSPERO was interrogated to confirm no recent/ongoing systematic reviews.

Eligible studies were observational (cohort, case-control and nested case-control; including letters or abstracts), using mNGS to characterise the virome in any clinical specimen in children aged \(\leq 18\) years who developed IA and/or T1D. Age restriction was imposed because IA often develops in childhood, suggesting viruses exert influence early in life. IA was defined as persistence of one or more autoantibodies against T1D-associated autoantigens (IAA, GADA, IA2A, ICA and ZnT8A) in \(\geq 2\) time-separated consecutive samples. Transplacental autoantibodies were excluded, defined as transient presence of the same autoantibody in a child \(\leq 18\) months and his/her mother. T1D was defined using American Diabetes Association criteria.\(^33\) Eligible studies were categorised into two groups based on the outcome: IA or T1D. Data were extracted on vertebrate-infecting viruses only, excluding studies that only analysed bacteriophage.

Two reviewers (Clare L. Faulkner and Yi Xuan Luo) screened titles and abstracts of identified studies (Figure 1) and then analysed shortlisted studies in full text for eligibility. In instances of uncertainty \((n = 2)\), an independent advisor (Ki Wook Kim) was consulted to reach consensus decision. Case reports/series, uncontrolled studies, reviews and animal studies were omitted based on exclusion criteria.
2.2 Data analysis

Data extracted included publication authors, year and geographical location; study design; study participants; number of cases/controls; age; level of pre-existing IA/T1D risk (human leukocyte antigen [HLA] genotype and family history); sample type, number and collection protocols; virus detection method and positivity threshold; rates of virus positivity in cases/controls; examined outcome (IA or T1D); measures of effect and funding. Original authors were contacted for insufficient or missing published data (n = 4).

Two reviewers (Clare L. Faulkner and Yi Xuan Luo) independently assessed the quality of included studies using the National Health and Medical Research Council (NHMRC) level of evidence scale and the Newcastle–Ottawa Quality Assessment Scale (NOS), as recommended by Cochrane collaboration. The NHMRC scale grades the study design according to a defined research hierarchy. The NOS evaluates three areas: selection, comparability and exposure; out of nine points, greater than six indicates good methods. Our chosen comparability controls were age and sampling time, and two critical factors likely to impact the prevalence of viruses.

We calculated ORs with 95% CIs and p-values for viruses present in children with IA or T1D versus controls from the extracted data using the Mantel–Haenszel method. Virus positivity was defined as virus material present in ≥1 study sample as detected by mNGS. Analysis was performed both for the number of case/control individuals and the number of case/control samples positive for virus. We used both fixed- and random-effects models; only results from random-effects models are presented due to heterogeneity of study populations. Statistical heterogeneity was explored using Cochrane's Q Test ($\chi^2$) and the $I^2$ statistic, which indicate the proportion of variance of the summary effect attributable to between-study heterogeneity. A $p < 0.10$ was considered a statistically significant heterogeneity, while $I^2 \leq 25\%$ and $\geq 75\%$ were deemed low and high heterogeneity, respectively. Subgroup analyses were performed for geographical location, stool versus plasma, consecutive virus shedding and studies using comparable detection thresholds, and pooled ORs were calculated. Sensitivity and influence analyses were conducted by the study size. Data analysis was completed in Review Manager, Version 5.4 (Cochrane Collaboration), with significance $p \leq 0.05$. 

FIGURE 1 Flow diagram of study selection
<table>
<thead>
<tr>
<th>Study</th>
<th>Country</th>
<th>Cases</th>
<th>Cases/controls</th>
<th>Autoantibodies measured</th>
<th>Age (years)</th>
<th>Control matching</th>
<th>Sample collection protocol</th>
<th>Total samples (cases/controls)</th>
<th>Virus sequencing; detection threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cinek et al.</td>
<td>Finland</td>
<td>Children who seroconverted less than 2 years old from DIPP</td>
<td>18/18</td>
<td>≥2 IAA, GADA, ICA, IA2A, or ZnT8A</td>
<td>0–2.5</td>
<td>Date/place of birth, HLA genotype, gender</td>
<td>Stool 3, 6 and 9 months before IA onset</td>
<td>92 (46/46)</td>
<td>mNGS only; threshold 50p100K</td>
</tr>
<tr>
<td>Hippich et al.</td>
<td>Germany</td>
<td>Ab+ children from BABYDIET</td>
<td>20/20</td>
<td>≥1 IAA, GADA, IA2A or ZnT8A</td>
<td>0–1.2</td>
<td>Age</td>
<td>PBMCs 3 Monthly from age 3 months</td>
<td>102 (51/51)</td>
<td>VirCapSeq-VERT and mNGS; threshold not stated</td>
</tr>
<tr>
<td>Kim et al.</td>
<td>Australia</td>
<td>Ab+ children from VIGR</td>
<td>20/20 (Stool study)</td>
<td>≥1 IAA, GADA or IA2A</td>
<td>5.7 ± 3.7</td>
<td>Age, gender</td>
<td>Stool At seroconversion and/or within 15 ± 6 months prior</td>
<td>64 (32/32)</td>
<td>VirCapSeq-VERT and mNGS; two thresholds: (1) 100 viral reads matched at species level and (2) 50p100K</td>
</tr>
<tr>
<td>Kramná et al.</td>
<td>Finland</td>
<td>Children who seroconverted less than 2 years old from DIPP</td>
<td>19/19</td>
<td>≥2 IAA, GADA, ICA, IA2A or ZnT8A</td>
<td>0–2</td>
<td>Date/place of birth, HLA genotype, gender</td>
<td>Stool 3, 6 and 9 months before IA onset</td>
<td>96 (48/48)</td>
<td>mNGS and retesting with PCR; threshold 50p100K</td>
</tr>
</tbody>
</table>

**TABLE 1** Summary of studies investigating the virome and IA
<table>
<thead>
<tr>
<th>Study</th>
<th>Country</th>
<th>Cases</th>
<th>Autoantibodies measured</th>
<th>Age (years)</th>
<th>Control matching</th>
<th>Sample collection protocol</th>
<th>Total samples (cases/controls)</th>
<th>Virus sequencing; detection threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vehik et al.</td>
<td>United States, Finland, Germany, Sweden</td>
<td>Ab+ children from TEDDY</td>
<td>≥1 IAA, GADA or IA2A</td>
<td>0-10</td>
<td>Age, clinical centre, gender, T1D family history</td>
<td>Monthly from age 3-48 months, quarterly thereafter; mean samples per subject 9</td>
<td>8,654 (4,327/4,327)</td>
<td>Culture to amplify low abundance viruses and mNGS; VirMAP aggregate bit score of 400 as threshold</td>
</tr>
<tr>
<td>Zhao et al.</td>
<td>Finland and Estonia</td>
<td>Ab+ children from DIABIMMUNE</td>
<td>≥1 IAA, GADA, IA2A, ICA or ZnT8</td>
<td>0-3</td>
<td>Age, gender, HLA genotype, birth delivery method, country</td>
<td>Monthly from 0 to 3 years; sequential samples analysed</td>
<td>220 (114/106)</td>
<td>mNGS; threshold not stated</td>
</tr>
<tr>
<td>Lee et al.</td>
<td>USA, Finland, Germany, Sweden</td>
<td>Ab+ children with rapid-onset T1D from TEDDY</td>
<td>≥1 IAA, GADA or IA2A</td>
<td>0-3</td>
<td>Age, clinical centre, T1D family history</td>
<td>Last Ab–negative visit and first Ab + seroconversion visit</td>
<td>56 (28/28)</td>
<td>mNGS; threshold not stated</td>
</tr>
</tbody>
</table>

Abbreviations: 50p100K, 50 viral reads per 100,000 raw reads; Ab–, autoantibody negative; Ab+, autoantibody positive; DIPP, Type 1 Diabetes Prediction and Prevention; FDR, first-degree relative; GADA, glutamic-acid decarboxylase autoantibodies; HLA, human leukocyte antigen; IA, islet autoimmunity; IA2A, tyrosine phosphatase-like insulinoma antigen 2 autoantibodies; IAA, islet autoantibodies; ICA, islet cell autoantibodies; mNGS, metagenomic next-generation sequencing; PBMC, peripheral blood mononuclear cell; T1D, type 1 diabetes; TEDDY, The Environmental Determinants of Diabetes in the Young; VIGR, Australian Viruses in the Genetically at Risk; VirCapSeq-VERT, Virome Capture Sequencing Platform for Vertebrate Viruses; ZnT8A, β-cell-specific zinc transporter 8 autoantibodies.

aData are reported as range, or mean ± SD.
TABLE 2 Summary of studies investigating the virome and T1D

<table>
<thead>
<tr>
<th>Study</th>
<th>Country</th>
<th>Cases/controls</th>
<th>Design/eligibility</th>
<th>Age (years)</th>
<th>Controls</th>
<th>Sample collection protocol</th>
<th>Total samples (cases/controls)</th>
<th>Virus sequencing and detection threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cinek et al.</td>
<td>Azerbaijan, Jordan,</td>
<td>73/105</td>
<td>Case control; patients with newly diagnosed T1D</td>
<td>&lt;18</td>
<td>Matched for age, place of residence</td>
<td>One stool sample collected shortly after T1D diagnosis</td>
<td>177 (73/104)</td>
<td>mNGS and specific PCR for EV, parechovirus, adenovirus, bocavirus, norovirus, sapovirus; threshold not stated</td>
</tr>
<tr>
<td></td>
<td>Nigeria, Sudan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vehik et al.</td>
<td>USA, Finland, Germany,</td>
<td>112/112</td>
<td>Nested-case control; high-risk HLA genotypes a</td>
<td>0–10</td>
<td>Matched for age, clinical centre, gender, T1D family history</td>
<td>Stool samples collected monthly from age 3 to 48 months, quarterly thereafter</td>
<td>3,380 (1,690/1,690)</td>
<td>Culture to amplify low abundance viruses and mNGS; VirMAP aggregate bit score of 400 as threshold</td>
</tr>
<tr>
<td></td>
<td>Sweden</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: EV, enterovirus; HLA, human leukocyte antigen; mNGS, metagenomic next-generation sequencing; PCR, polymerase chain reaction; T1D, type 1 diabetes.

aHigh-risk HLA genotypes include DR3/4, DR4/4, DR4/8 and DR3/3.

Our systematic review is reported using meta-analysis of observational studies in epidemiology and Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines.

3 | RESULTS

The search returned 778 publications (51 duplicates), leaving 727 articles for review. Title and abstract screening identified 12 publications for full-text review. Four were excluded: three were repeat data sets; one used targeted polymerase chain reaction (PCR) rather than pre-specified mNGS; eight were included—one abstract, two letters and five articles (Figure 1). Two publications contained two study groups that were analysed separately, giving a total 10 studies with 1,425 participants (695 cases, 730 controls). Nine were nested case–control studies using samples collected within prospective birth cohorts, eight investigated IA (510 cases, 513 controls) and two investigated T1D (185 cases, 217 controls; Tables 1 and 2). One IA study with insufficient data was excluded from meta-analysis.

3.1 | Study characteristics

Six studies defined IA as positivity for ≥1 T1D-associated autoantibody; two defined ≥2 autoantibodies (Table 2). All IA studies, except two, required persistent autoantibody positivity across consecutive visits. All IA and T1D nested case–control studies selected participants from within the same prospective cohort. Most prospective cohorts recruited participants with high-risk HLA genotypes (DR3/4, DR4/4, DR4/8 and DR3/3), except one that recruited children with ≥1 first-degree relative (FDR) with T1D, and one that required both the criteria. Most studies analysed children less than 6 years; two investigated older children ≤10 and ≤18 years. Most studies included <50 participants. Study characteristics are summarised in Tables 1 and 2.

Seven studies examined the gut virome by sequencing virus-enriched stool; two investigated plasma and one examined peripheral blood mononuclear cells (PBMCs). Alongside mNGS, four studies utilised specific PCR for common viruses.

Three studies employed Virome Capture Sequencing Platform for Vertebrate Viruses (VirCapSeq-VERT) to enhance sensitivity for vertebrate-infecting viruses. Two studies cultivated stool in virus-susceptible cells to amplify low-abundance EV or other common viruses.

Viruses commonly reported in the IA group included EV, bocaparvovirus, anelloviruses, parechovirus, rotavirus, sapovirus, cardiovirus and mastadenovirus. Norovirus, circovirus, mamastrovirus, kobuvirus, picobirnavirus, erthroparvovirus and roseolovirus positivity were only reported in two studies, and eight other viruses were reported each in only one study. Viruses commonly reported in the two T1D studies were EV, parechovirus, bocaparvovirus, anelloviruses, sapovirus, cardiovirus, mastadenovirus, norovirus and mamastrovirus. Kobuvirus and circovirus were only reported in one study.

3.2 | Quality of evidence

The NOS scores were ≥8 (Table 3), indicating good methodological quality overall. Of the 10 studies, only 1 IA and 1 T1D study adjusted for potential confounders, including HLA genotype, lifestyle factors, demographic factors and factors related to the child.
TABLE 3  Quality of evidence in observational studies investigating the virome and islet autoimmunity or type 1 diabetes

<table>
<thead>
<tr>
<th>Study</th>
<th>NHMRC level of evidence</th>
<th>Newcastle–Ottawa Scale Score</th>
<th>Cases and controls matched?</th>
<th>Details of virome sequencing Method given?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Selection</td>
<td>Comparability</td>
<td>Exposure</td>
<td>Total/Nine</td>
</tr>
<tr>
<td>Cinek et al48</td>
<td>II</td>
<td>●●●●</td>
<td>●●</td>
<td>9</td>
</tr>
<tr>
<td>Cinek et al41</td>
<td>III-3</td>
<td>○●●●</td>
<td>●●</td>
<td>8</td>
</tr>
<tr>
<td>Hippich et al.43</td>
<td>II</td>
<td>●●●●</td>
<td>●●</td>
<td>8</td>
</tr>
<tr>
<td>Kim et al.47</td>
<td>II</td>
<td>●●●●</td>
<td>●●</td>
<td>9</td>
</tr>
<tr>
<td>Kramná et al.44</td>
<td>II</td>
<td>●●●●</td>
<td>●●</td>
<td>9</td>
</tr>
<tr>
<td>Lee et al.45</td>
<td>II</td>
<td>●●●●</td>
<td>●●</td>
<td>9</td>
</tr>
<tr>
<td>Vehik et al.42</td>
<td>II</td>
<td>●●●●</td>
<td>●●</td>
<td>9</td>
</tr>
<tr>
<td>Zhao et al.46</td>
<td>II</td>
<td>●●●●</td>
<td>●●</td>
<td>9</td>
</tr>
</tbody>
</table>

Note: ● = 1 point; N/A, not available.

*II, nested case-control study; III-3, case-control study.

Meta-analyses for specific viruses were conducted where proportions of positive case-control individuals or samples were reported in ≥1 study. Six studies found a significant association between the number of EV-positive samples and IA (1.13, 1.00–1.28, p = 0.05). For stool versus plasma subgroup analysis, the pooled ORs were 1.14 (1.00–1.29, p = 0.05) and 0.80 (0.24–2.73, p = 0.73), respectively (Figure 3). There was no association between EV-positive individuals and IA (1.13, 0.86–1.48, p = 0.37). For stool versus plasma subgroup analysis, pooled ORs were 1.15 (0.87–1.51, p = 0.32) and 0.80 (0.23–2.77, p = 0.72), respectively. There were minimal differences in effect sizes with sensitivity analysis.

3.3 | Islet autoimmunity

Seven studies investigated vertebrate-infecting viruses and IA. No significant heterogeneity was observed, unless stated. Due to insufficient sample data, positivity for any virus was only analysed at the individual level, with no difference between cases and controls, pooled OR 1.03 (95% CI 0.67–1.60, p = 0.89). Stool versus plasma subgroup analysis gave pooled ORs 1.20 (0.71–2.04, p = 0.49) and 0.66 (0.29–1.51, p = 0.33), respectively (Figure 2); the PBMC study43 was excluded. For European versus non-European (Australian) subgroup analyses, pooled ORs were 1.14 (0.66–1.96, p = 0.65) and 0.87 (0.41–1.82, p = 0.71), respectively.
versus association no summary χ odds sequencing. was (EV ‐ rates p IA. enterovirus based = 1.20; versus islet number children 95% IA samples metagenomic confidence EV B) between for by ‐ of results with next generation 4 (OR, B and found and of stool as for IA) All positive samples virus detected B FIGURE on ‐ FIGURE 1.00–1.29, between Individual (odds EV based no enterovirus on summary All with stool positive number positive association in IA with for samples rates sequencing. virus 1.14; positivity detected children interval ratio p EV 3 IA, 0.05; of plasma samples and next results for odds of χ number An 8 14 heterogeneity generation 1.61 between EV p pooled consecutive EVB = 1.20 (0.99–1.42, p = 0.04; Figure 4), but not individuals (0.99, 0.74–1.32, p = 0.94). There were no associations between IA and EV-A-positive samples or individuals, pooled ORs 1.61 (0.43–5.94, p = 0.48; significant heterogeneity $\chi^2 = 2.81, p = 0.09, I^2 = 64\%$) and 1.12 (0.84–1.50, p = 0.42), respectively.

Three stool studies reported consecutive EV shedding (≥2 sequential samples positive) and IA, pooled OR 1.55 (1.09–2.20, p = 0.01; Figure 5). One study reported consecutive EV-A/EV-B shedding: IA was associated with consecutive EV-B (2.46, 1.46–4.16, p = 0.0007), but not EV-A (1.19, 0.74–1.92, p = 0.47).

Four studies reported parechovirus positivity in stool. There was no association between the number of individuals positive for parechovirus and IA, pooled OR 0.83 (0.63–1.10, p = 0.20). However, for parechovirus-positive samples and IA (0.66, 0.32–1.35; p = 0.25), there was significant heterogeneity ($\chi^2 = 3.88; p = 0.009; I^2 = 74\%$). In influence analysis, the removal of the largest study outlier (>8,000 samples) strengthened the magnitude of association, pooled OR 0.44 (0.23–0.81, p = 0.008) and low heterogeneity ($\chi^2 = 1.10, p = 0.33, I^2 = 9\%$). There was minimal difference in effect size for the number of individuals positive for parechovirus.

Meta-analyses at the individual and sample level for rotavirus, bocaparvovirus, anelloviruses, sapovirus, norovirus, cardiovirus, circovirus, mamastrovirus, mastadenovirus, kobuvirus, picobirnavirus, erythroparvovirus and roseolovirus showed no associations with IA (Table S1). There were minimal differences in effect sizes with sensitivity and influence analyses. Viruses reported in only one study were precluded from the meta-analysis. One study conducted strain-specific analysis for mastadenovirus (human mastadenovirus A, C, F) and found association between the number of samples with positive human mastadenovirus F and IA, OR 1.33 (1.08–1.54, p = 0.007).

### Table 3

<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>IA Cases Total</th>
<th>IA Controls Total</th>
<th>Total</th>
<th>Odds Ratio M-H, Random, 95% CI Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Krammá et al. 2015</td>
<td>0 48</td>
<td>1 48</td>
<td>0.1%</td>
<td>0.33 [0.01, 8.22] 2015</td>
</tr>
<tr>
<td>Zhao et al. 2017</td>
<td>10 114</td>
<td>5 106</td>
<td>1.2%</td>
<td>1.94 [0.64, 5.88] 2017</td>
</tr>
<tr>
<td>Vahik et al. 2019 (Outcome: IA)</td>
<td>590 4327</td>
<td>530 4327</td>
<td>96.7%</td>
<td>1.13 [1.00, 1.28] 2019</td>
</tr>
<tr>
<td>Kim et al. 2019 (Stool Virome)</td>
<td>5 32</td>
<td>6 32</td>
<td>0.8%</td>
<td>1.00 [0.26, 3.86] 2019</td>
</tr>
<tr>
<td>Subtotal (95% CI)</td>
<td>4521</td>
<td>4513</td>
<td>95.0%</td>
<td>1.14 [1.00, 1.29]</td>
</tr>
<tr>
<td>Total events</td>
<td>605 541</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heterogeneity: Tau² = 0.00; Chi² = 1.51, df = 3 (p = 0.68); I² = 0% Test for overall effect: Z = 2.00 (p = 0.05)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 4

<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>IA Cases Total</th>
<th>IA Controls Total</th>
<th>Total</th>
<th>Odds Ratio M-H, Random, 95% CI Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lee et al. 2013</td>
<td>1 28</td>
<td>0 28</td>
<td>0.1%</td>
<td>3.11 [0.12, 79.64] 2013</td>
</tr>
<tr>
<td>Kim et al. 2019 (Plasma Virome)</td>
<td>4 59</td>
<td>6 59</td>
<td>0.9%</td>
<td>0.64 [0.17, 2.41] 2019</td>
</tr>
<tr>
<td>Subtotal (95% CI)</td>
<td>87</td>
<td>87</td>
<td>1.0%</td>
<td>0.80 [0.24, 2.73]</td>
</tr>
<tr>
<td>Total (95% CI)</td>
<td>4608</td>
<td>4600</td>
<td>100.0%</td>
<td>1.13 [1.00, 1.28]</td>
</tr>
<tr>
<td>Total events</td>
<td>610 547</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heterogeneity: Tau² = 0.00; Chi² = 2.60, df = 5 (p = 0.78); I² = 0% Test for overall effect: Z = 1.96 (p = 0.05) Test for subgroup differences: Chi² = 0.30, df = 1 (p = 0.58), I² = 0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 3** Individual and summary odds ratios for number of samples positive for enterovirus (EV) in children with islet autoimmunity (IA) versus no IA, with stool versus plasma subgroup analysis. All results based on rates of virus positivity as detected by metagenomic next-generation sequencing. An association was found between childhood IA and number of stool samples positive for EV (odds ratio 1.14; 95% confidence interval 1.00–1.29, p = 0.05; heterogeneity $\chi^2 = 0.50, p = 0.68, I^2 = 0\%$), but not the number of plasma samples positive for EV.

**Figure 4** Individual and summary odds ratios (ORs) for number of stool samples positive for enterovirus B (EV-B) in children with islet autoimmunity (IA) versus no IA. All results based on rates of virus positivity as detected by metagenomic next-generation sequencing. An association was found between number of stool samples positive for EV-B and IA (OR, 1.20; 95% confidence interval 1.01–1.42, p = 0.04; heterogeneity $\chi^2 = 0.03, p = 0.86, I^2 = 0\%$).
Meta-analysis of consecutive shedding of parechovirus, bocaparvovirus, anelloviruses and picornavirus found no associations with IA (Table S2). Meta-analysis of viruses reported in the two studies\textsuperscript{44,47} applying the same positivity threshold of 50 viral reads per 100,000 raw reads (50p100K) found no associations between any virus and IA, including for EV, parechovirus, anelloviruses, bocaparvovirus and sapovirus (Table S3).

Study quality subgroup analysis was excluded as NOS scores were $\geq$8. HLA subgroup analysis was excluded as no studies stratified virus positivity by genotype, and all but two\textsuperscript{47} recruited only high-risk genotypes. Geographical location subgroup analysis for individual viruses was excluded as only two studies were non-European\textsuperscript{47} and multicentre studies insufﬁciently compared study populations.

Only one study\textsuperscript{47} examined differential abundance of viruses in the gut of children with IA versus controls, precluding meta-analysis. It found 129 viruses with more than twofold difference in cases versus controls ($p = 0.02$). Notably, human mastadenovirus F, astrovirus, human adenovirus 41, coxsackievirus A2 (CVA2), enteric cytopathic human orphan virus 30, coxsackievirus B3 and human parechovirus were more abundant in cases, while saffold virus, norovirus and rotavirus A were more abundant in controls. Every differentially abundant rotavirus A (CVA2, 5, 6, 8, 14) was more abundant in cases ($p < 0.00001$). Additionally, one study\textsuperscript{46} measured intestinal viral alpha and beta diversity in children with IA versus controls, with the gut viromes of cases signiﬁcantly less diverse (with lower interpersonal variation) compared to controls ($p < 0.001$).

One analysis of very young children <6 months\textsuperscript{42} demonstrated the association between early-life human mastadenovirus C infection and lower IA risk (0.55, 0.38–0.80, $p = 0.001$).

3.4 | Type 1 diabetes

Two studies\textsuperscript{41,42} investigated gut vertebrate-infecting viruses and T1D, with no association between positivity for any virus and T1D, pooled OR 0.94 (0.54–1.64, $p = 0.83$; and no heterogeneity $\chi^2 = 0.21$, $p = 0.65$, $I^2 = 0$).

Meta-analyses for EV, parechovirus, cardioivirus, norovirus, sapovirus, mastadenovirus, bocaparvovirus, mamastrovirus and anelloviruses were not significant for the number of positive individuals (Table S4). Strain-specific analysis for mastadenovirus produced no signiﬁcant effect sizes, including for human mastadenovirus A (OR 0.92, 0.51–1.65, $p = 0.78$) and human mastadenovirus F (OR 0.72, 0.42–1.24, $p = 0.24$), with no heterogeneity between studies. One study\textsuperscript{42} reported virus-positive sample numbers, precluding meta-analysis. One study\textsuperscript{42} analysed EV subtypes, reporting an association between T1D protection and number of EV-B positive samples (0.73, 0.53–0.99, $p = 0.05$), but not individuals (0.69, 0.41–1.18, $p = 0.18$); EV-A was not associated with T1D. Limited studies precluded subgroup and sensitivity analyses.

4 | DISCUSSION

This systematic review of 10 observational studies, involving 695 cases and 730 controls, demonstrated associations between virome composition in children ≤18 years and development of IA, but not T1D. This suggests early virome changes may inﬂuence initiation of IA, but not progression to T1D. There was a weak association between the number of stool samples positive for EV or EV-B and IA, with approximately 1.2 times the odds of EV or EV-B positivity in children who developed IA versus controls; ORs 1.14 (1.00–1.29) and 1.20 (1.01–1.42), respectively. There was 1.5 times the odds of consecutive EV shedding in stool of children with IA versus controls; OR 1.55 (1.09–2.20). Only one study measured consecutive shedding of EV serotypes,\textsuperscript{42} demonstrating signiﬁcant association between consecutive EV-B positivity and IA (2.46; 1.46–4.16), but not EV-A (1.19, 0.74–1.92). Influence analysis, removing the largest outlier study, demonstrated half the odds of parechovirus shedding in stool for children with IA versus controls (0.44, 0.23–0.81). Other viruses were not associated with IA at the individual or sample level.

These data suggest that specific gut vertebrate-infecting viruses present in the gut virome, rather than the presence of any virus, inﬂuence IA risk. Research repeatedly reports associations between EV infection and IA initiation, supported by EV RNA in stools\textsuperscript{40} and EV RNA/antibodies in sera\textsuperscript{16,40,49,50}. Our results support clinical studies\textsuperscript{51,52} and pancreatic tropism studies\textsuperscript{20,53,54} favouring EV-B as a candidate virus in IA susceptibility. In contrast, studies of other candidate viruses remain inconclusive. For example, parechovirus
may confer IA protection; however, no associations with IA/T1D have previously been reported.  

We demonstrated an association between IA and the number of virus-positive samples, but not individuals. Thus, when an individual has more than one positive sample, their IA risk is amplified. This may relate to viral persistence, protracted infection or increased exposure through reinfection. In particular, IA was associated with consecutive or prolonged EV/EV-B shedding. Consecutive shedding is a strong indicator of persistent infection. Viral persistence is suggested to play a critical role in the development of autoimmunity through ongoing aberrant presentation of antigens to the immune system, production of inflammatory cytokines and induction of Endoplasmic Reticulum stress. The gut mucosa may be a potential viral reservoir for sustained pancreatic infection with multiple virus-positive stool samples a marker of persistent gut infection. Consecutive shedding may also indicate defective or dysregulated innate immune defence, which increases autoimmune propensity. Longitudinal virome studies are therefore essential in tracking virus infections over time. However, our ability in this review to distinguish between persistence of the same viral strain or reinfection in consecutively positive patient samples was limited by most studies reporting viruses detected at the genus level and intermittent sampling across studies.

The one study that conducted differential abundance analysis found 129 viruses with a 2-fold difference in abundance in the gut of IA cases versus controls. This suggests IA risk is closely linked to viral load of a variety of viruses. Higher viral titre facilitates greater replication, pancreatic transmission, persistence, cellular stress and establishment of an immunogenic environment. Future mNGS studies in larger cohorts with more timepoints preceding IA or T1D are required to elucidate IA/T1D-associated vertebrate-infecting viruses and compare differential abundance across the breadth of potentially diabetogenic viruses. Additionally, further research investigating virome composition across defined early-life stages is required to determine time points where viruses exert greatest influence. Only one study measured ‘virome diversity’, finding lower diversity in children with IA. Further development and standardisation of these diversity measures are required to facilitate comparability between studies and greater understanding of association between virome composition and IA/T1D risk.

4.1 Strengths and weaknesses

To minimise bias, we implemented pre-defined eligibility criteria, screening by independent reviewers, no language restrictions and sources beyond indexed databases. Random-effects models may have provided more conservative effect estimates by accounting for study population heterogeneity and generating wider confidence intervals. Limiting investigation to children ≤18 years may have skewed results due to high rates of childhood background infection. However, there are no adult IA/T1D virome studies. We included studies conducted globally to minimise geographical bias related to infection rates. However, eight studies were European, where T1D incidence is the highest. This precluded country subgroup analysis in most meta-analyses. All studies recruited infants with high-risk HLA or an affected FDR, potentially reducing generalisability.

Our findings have limitations. Only two studies examined T1D, limiting the analysis of associations between viruses and progression to T1D beyond IA initiation. Of the two studies that analysed T1D, one analysed prospectively collected samples and one analysed samples collected at/after T1D diagnosis. This may limit comparability due to potential differences when examining the virome after, rather than before, diagnosis of the study outcome. IA was predominantly defined as ≥1 autoantibody, despite single autoantibody conferring lower lifetime T1D risk versus multiple antibodies. However, our stratification of results by autoantibody number was precluded by insufficient data. Most studies matched for HLA genotype, but HLA subgroup analysis could not be conducted, despite HLA predicting IA/T1D risk and potentially influencing virus-induced pathology, immune dysregulation or susceptibility to viral infection. However, studies that have explored the association with HLA and EV infection specifically have reported inconsistent results, finding varying infection prevalence in individuals with different HLA genotypes and no association. Thus, future studies of infants with a range of low- to high-risk genotypes are required, such as Environmental Determinants of Islet Autoimmunity (ENDIA). Many studies did not account for other potential environmental risk factors, such as anthropometry, diet, vitamin D, omega-3 fatty acids, birth delivery route and/or breastfeeding, all of which can influence the viral presence and IA/T1D risk. However, controlling for all potential confounders in small case-control studies remains challenging, with only the two largest studies included in this review reporting adjustment for a number of potential confounders.

Studies applied different positivity thresholds, including 50p100K and VirMAP aggregate bit score; several studies did not report thresholds. Thresholds maintain sensitivity while minimising false positives and reducing low-level cross-sample contamination risk, but variability limits comparability, potentially introducing errors into meta-analysis and its interpretation. Study method heterogeneity was also present, including mixed use of culture, PCR and various mNGS platforms. For example, TEDDY’s use of culture to amplify low-abundance viruses impeded comparative analysis of viral load and increased bias to detecting certain viruses above threshold.

Despite advances in metagenomics, virome analysis remains challenging, with rapid viral evolution complicating sequencing and impeding novel virus mapping by bioinformatic databases. Indeed in most viral mNGS datasets, more than 50% of sequences exhibit no detectable sequence similarity to known reference sequences, contributing to the viral ‘dark matter’. It is plausible that these may include highly divergent or completely novel viruses that have yet to be discovered. Small genomes and low abundance of vertebrate-infecting viruses in human samples hampers detection, with high background interference from other genetic material.
Thus, effective viral enrichment is necessary, such as enzymatic digestion of non-viral nucleic acids or size exclusion of non-viral components via filtration. Three studies employed VirCapSeq-VERT and demonstrated enhanced sensitivity for identifying a broader range of vertebrate-infecting viruses. VirCapSeq-VERT uses approximately two million probes targeting genomes of all known vertebrate-infecting viruses, increasing the sensitivity of viral sequence detection by up to 10,000-fold compared to standard mNGS. Wider application of VirCapSeq-VERT or other similar pan-viral enrichment sequencing platforms will significantly enhance the reproducibility and robustness of future virome studies.

Positivity for viral nucleic acid is a marker of infection, not proof, as viruses may pass through the gut without productive infection, as with plant viruses and diet-derived viruses. Similarly, viral shedding in stool/plasma cannot directly evidence pancreatic infection. Additionally, periodic sample collection precludes determination of first virus exposure, differentiation between persistence or re-infection over time and definition of precise temporal associations between infection and IA/T1D onset. However, more frequent sampling may not be sustainable over long follow-up in prospective cohort studies and may still miss some acute infections with a very narrow window for detection.

Studies sampled various body sites through stool, plasma and PBMCs. Direct comparison of across sites is difficult, requiring careful consideration of where viruses replicate. For example, EVs and mastadenoviruses preferentially infect and replicate at mucosal surfaces and gut viral shedding in stool persists longer, resulting in higher positivity compared to short viraemic periods in plasma/PBMCs.

Finally, this review did not examine bacteriophage to limit scope and adopt the precedent in other studies of focusing on only one class of virus to maximise detection sensitivity.

Overall quality of included studies was high (NOS scores ≥8). All studies except one matched for ≥3 factors. Meta-analysis demonstrated little significant heterogeneity; however, results must be interpreted cautiously given χ² and I² limitations in detecting true heterogeneity. Studies were small: six had <50 participants, potentially causing small study effects. However, longitudinal sampling increased statistical power for detecting differences in the virome between cases and controls. Thus, we demonstrated the importance of comparing both number of virus-positive individuals and samples. One study was significantly larger, with high weighting in meta-analysis, potentially skewing our results. Influence analyses, removing the smallest and largest outlier studies in turn, demonstrated insufficient studies of TEDDY scale. The ongoing ENDIA study will contribute significantly as a large, nationwide observational, prospective cohort of 1,500 children followed from pregnancy through early life.

4.2 | Future research

Despite limitations of targeted viral detection in IA/T1D pathogenesis studies, there remains a paucity of large, unbiased virome studies. Our findings must be validated in future studies that (1) include >200 participants and frequent longitudinal sampling preceding IA/T1D onset to improve statistical power and counter small-study effects; (2) include a wider range of HLA genotypes to consider viral associations with IA/T1D in the context of genetic risk; (3) incorporate multicentre data to reduce geographical bias; (4) employ sensitive enrichment and comprehensive sequencing platforms; (5) integrate differential abundance analysis of viral load; and (6) sample various body sites to characterise viral strains and account for niche variations in viral abundance. Future studies should also include virome analysis during pregnancy to explore the role of antenatal and congenital infections in offspring IA/T1D development.

ACKNOWLEDGEMENTS

This work was supported by the National Health and Medical Research Council Practitioner fellowship APP1136735 (to Maria E. Craig), the Juvenile Diabetes Research Foundation International Postdoctoral Fellowship 3-PDF-2020-940-A-N and the Australian Diabetes Society Lindsey Baudinet Award (to Ki Wook Kim). The authors would like to thank Professors Ondrej Cinek and Kendra Vehik for providing supplementary raw data for inclusion in this systematic review.

CONFLICT OF INTEREST

The author declares that there is no conflict of interests.

AUTHOR CONTRIBUTIONS

Maria E. Craig and Ki Wook Kim designed the study and led the study group. Clare L. Faulkner performed the literature search, extracted, analysed and interpreted the data and wrote the manuscript. Yi Xuan Luo independently confirmed the database search and study quality. Ki Wook Kim acted as consultant in instances of uncertainty. Yi Xuan Luo, Sonia Isaacs, William D. Rawlinson, Maria E. Craig and Ki Wook Kim helped interpret the data. All authors were involved in the critical revision of this manuscript. Ki Wook Kim is the guarantor of this study.

ORCID

William D. Rawlinson https://orcid.org/0000-0001-9238-0546
Maria E. Craig https://orcid.org/0000-0001-6004-576X
Ki Wook Kim https://orcid.org/0000-0001-9579-6408

REFERENCES


**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of this article.

Early-life factors contributing to type 1 diabetes

Maria E. Craig 1,2,3 • Ki Wook Kim 1,4 • Sonia R. Isaacs 1,4 • Megan A. Penno 5,6 • Emma E. Hamilton-Williams 7 • Jennifer J. Couper 5,6 • William D. Rawlinson 1,4,8

Received: 22 March 2019 / Accepted: 21 May 2019 © Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract
The incidence of type 1 diabetes has increased since the mid-twentieth century at a rate that is too rapid to be attributed to genetic predisposition alone. While the disease can occur at any age, mounting evidence from longitudinal cohort studies of at-risk children indicate that type 1 diabetes associated autoantibodies can be present from the first year of life, and that those who develop type 1 diabetes at a young age have a more aggressive form of the disease. This corroborates the hypothesis that environmental exposures in early life contribute to type 1 diabetes risk, whether related to maternal influences on the fetus during pregnancy, neonatal factors or later effects during infancy and early childhood. Studies to date show a range of environmental triggers acting at different time points, suggesting a multifactorial model of genetic and environmental factors in the pathogenesis of type 1 diabetes, which integrally involves a dialogue between the immune system and pancreatic beta cells. For example, breastfeeding may have a weak protective effect on type 1 diabetes risk, while use of an extensively hydrolysed formula does not. Additionally, exposure to being overweight pre-conception, both in utero and postnatally, is associated with increased risk of type 1 diabetes. Epidemiological, clinical and pathological studies in humans support a role for viral infections, particularly enteroviruses, in type 1 diabetes, but definitive proof is lacking. The role of the early microbiome and its perturbations in islet autoimmunity and type 1 diabetes is the subject of investigation in ongoing cohort studies. Understanding the interactions between environmental exposures and the human genome and metagenome, particularly across ethnically diverse populations, will be critical for the development of future strategies for primary prevention of type 1 diabetes.

Keywords Child • Diet • Enterovirus • Gene–environment interaction • Life course development • Microbiome • Pregnancy • Review • Type 1 diabetes • Virus

Abbreviations
CVB Coxsackievirus group B
DAISY Diabetes Autoimmunity Study in the Young
DIPP Diabetes Prediction and Prevention
ENDIA Environmental Determinants of Islet Autoimmunity
ER Endoplasmic reticulum

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00125-019-4942-x) contains a slideset of the figures for download, which is available to authorised users.

Maria E. Craig
m.craig@unsw.edu.au

1 School of Women’s and Children’s Health, University of New South Wales Faculty of Medicine, Sydney, NSW, Australia
2 Institute of Endocrinology and Diabetes, Children’s Hospital at Westmead, Locked Bag 4001, Westmead, Sydney, NSW 2145, Australia
3 Discipline of Child and Adolescent Health, University of Sydney, Sydney, NSW, Australia
4 Virology Research Laboratory, Prince of Wales Hospital Randwick, Sydney, NSW, Australia
5 Robinson Research Institute, School of Paediatrics and Reproductive Health, University of Adelaide, Adelaide, SA, Australia
6 Department of Endocrinology and Diabetes, Women’s and Children’s Hospital, Adelaide, SA, Australia
7 University of Queensland Diamantina Institute, University of Queensland, Translational Research Institute, Brisbane, QLD, Australia
8 Serology and Virology Division, NSW Health Pathology, Prince of Wales Hospital, Sydney, NSW, Australia
Introduction

The accepted model that type 1 diabetes develops in an individual due to the interplay between genetic predisposition and environmental determinants has been challenged by the observation that the rising incidence of type 1 diabetes in individuals over 50 years of age is primarily associated with people who are not at increased genetic risk [1–3]. Hence, environmental determinants now appear to exert a greater influence on the risk of islet autoimmunity (IA) and type 1 diabetes than in the past. The large body of evidence to support this hypothesis is based on in vitro models, animal models, cohort studies, epidemiological analyses and in vivo observations. However, inconsistencies regarding the relative contribution of genes vs environment remain, which may be explained by the heterogeneity of type 1 diabetes across the lifespan and the globe, methodological differences in studies from which conclusions have been derived and the changing impact of factors over time. Moreover, there is a paucity of information on how the in utero milieu influences type 1 diabetes risk in the offspring. In the context of a changing permissive environment, we review the contribution of candidate early-life factors, from pregnancy through to early childhood, to the development of IA and progression to type 1 diabetes. The major cohort studies and trials referenced in this review are shown in Table 1 and factors associated with initiation of or protection from IA and progression to type 1 diabetes are summarised in the Text box.
The risk of type 1 diabetes is tenfold higher in children with an affected first-degree relative (FDR) compared with the general population. A substantial component of this risk (~50%) can be explained by specific HLA alleles [37]. The highest risk haplotypes are DRB1*03:01-DQA1*05:01-DQB1*02:01 (known as ‘DR3’) and DRB1*04-DQA1*03:01-DQB1*03:02 (known as ‘DR4’). In a pooled analysis of 5196 individuals with type 1 diabetes and 6359 control participants, individuals who were heterozygous for DR3/4 had an OR >16 for type 1 diabetes, while DR3/3 and DR4/4 homozygotes had ORs ~6 [37]. However, <10% of individuals with HLA-conferred susceptibility develop type 1 diabetes.

Genome-wide association studies elucidated a putative role for multiple non-HLA type 1 diabetes risk loci [38]. Incorporation of multiple loci into a genetic risk score (GRS) can predict >10% of risk for pre-symptomatic type 1 diabetes in children without an affected FDR [39]. In an ethnically diverse population from South-Eastern USA, younger age at type 1 diabetes diagnosis was associated with a higher prevalence of the DR3/4 diplotype and a higher GRS [40]. This suggests that the relative contribution of the environment may be greater in those who develop type 1 diabetes at an older age. Recently, the combination of 67 single nucleotide polymorphisms, HLA DR-DQ loci and their interactions into a revised GRS (GRS2) was highly discriminative for type 1 diabetes, particularly early-onset disease, with an area under the curve of 0.96 [41]. In the context of population screening for type 1 diabetes, which is now being undertaken through multiple programmes [42, 43], the GRS has substantial potential for prediction of type 1 diabetes in early life.

The application of a GRS has also provided further insights into the risk of IA and type 1 diabetes in childhood in familial vs non-familial type 1 diabetes. The Environmental Determinants of Diabetes in the Young (TEDDY) study screened infants for multiple type 1 diabetes-associated alleles and stratified genetic susceptibility into four groups based on presence of ‘high-risk’ DR4 allele subtypes and quartiles of GRS [27]. On comparing children with or without an FDR with type 1 diabetes, those in the highest risk group (high-risk DR4 allele subtypes and GRS <50th percentile) exhibited a markedly higher risk of IA relative to children without an FDR (9.2% vs 1.6%, respectively), suggesting that additional factors were enriched within families where more than one individual developed type 1 diabetes compared with the general population. The findings support the hypothesis that the contribution of genetic and environmental factors to the development of type 1 diabetes varies according to family history.

Table 1  Cohort studies and clinical trials investigating the relationship between early-life factors and IA and/or type 1 diabetes

<table>
<thead>
<tr>
<th>Cohort study/trial</th>
<th>Country/region</th>
<th>Study population</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australian BABYDIAB</td>
<td>Australia</td>
<td>Birth cohort: FDR with T1D</td>
<td>[4, 5]</td>
</tr>
<tr>
<td>BABYDIAB</td>
<td>Germany</td>
<td>Birth cohort: offspring of parents with T1D</td>
<td>[6]</td>
</tr>
<tr>
<td>BABYDIET</td>
<td>Germany</td>
<td>RCT: at least two FDRs with T1D, or one FDR with T1D and a high-risk HLA genotype</td>
<td>[7–9]</td>
</tr>
<tr>
<td>DAISY</td>
<td>CO, USA</td>
<td>Birth cohort: high-risk HLA genotypes or FDR with T1D</td>
<td>[10, 11]</td>
</tr>
<tr>
<td>DIABIMMUNE</td>
<td>Finland, Estonia, Russian Karelia</td>
<td>Birth cohort: high-risk HLA genotypes</td>
<td>[12, 13]</td>
</tr>
<tr>
<td>DIPP</td>
<td>Finland</td>
<td>Birth cohort: high-risk HLA genotypes</td>
<td>[14–18]</td>
</tr>
<tr>
<td>DNBC</td>
<td>Denmark</td>
<td>Birth cohort</td>
<td>[19, 20]</td>
</tr>
<tr>
<td>ENDIA</td>
<td>Australia</td>
<td>Pregnancy/birth cohort: FDR with T1D</td>
<td>[21, 22]</td>
</tr>
<tr>
<td>KVB claims data</td>
<td>Bavaria, Germany</td>
<td>Birth cohort</td>
<td>[23]</td>
</tr>
<tr>
<td>MIDIA</td>
<td>Norway</td>
<td>Birth cohort: high-risk HLA genotypes</td>
<td>[24]</td>
</tr>
<tr>
<td>MoBa</td>
<td>Norway</td>
<td>Birth cohort</td>
<td>[19, 20]</td>
</tr>
<tr>
<td>National Health Insurance Research database</td>
<td>Taiwan</td>
<td>Birth cohort</td>
<td>[25]</td>
</tr>
<tr>
<td>Tasmanian Infant Health Survey</td>
<td>TAS, Australia</td>
<td>Birth cohort</td>
<td>[26]</td>
</tr>
<tr>
<td>TEDDY</td>
<td>USA, Europe</td>
<td>Birth cohort: high-risk HLA genotypes or FDR with T1D</td>
<td>[27–33]</td>
</tr>
<tr>
<td>TRIGR</td>
<td>Europe, Canada, Australia</td>
<td>Primary prevention RCT from birth: FDR and high-risk HLA genotypes</td>
<td>[34, 35]</td>
</tr>
<tr>
<td>VIGR</td>
<td>Australia</td>
<td>Birth cohort: FDR with T1D</td>
<td>[36]</td>
</tr>
</tbody>
</table>

DNBC, Danish National Birth Cohort; KVB, Kassenärztliche Vereinigung Bayern; MIDIA, Environmental Triggers of Type 1 Diabetes; MoBa, Norwegian Mother and Child Cohort Study; T1D, type 1 diabetes
Dietary factors

Dietary factors, including early introduction of cow’s milk protein [14], overall dairy intake and early or late introduction of gluten [28], have long been implicated in the development of IA and type 1 diabetes, potentially through a mechanistic role; these dietary factors may act as antigenic triggers of autoimmunity or as co-factors in the context of gut infection and/or inflammation.

Maternal diet

The impact of maternal diet on the risk of type 1 diabetes has not been extensively assessed. In the Danish National Birth Cohort study, offspring of women with high gluten intake (>20 g/day) had double the risk of type 1 diabetes vs offspring of women with low gluten intake (<7 g/day) [44]. In the Diabetes Autoimmunity Study in the Young (DAISY), maternal intake of vitamin D via food (but not via supplements) was associated with a 63% decreased risk of IA in offspring [45]. Higher maternal vitamin D-binding protein levels and higher cord blood 25-hydroxyvitamin D [25(OH)D] was associated with a lower risk of type 1 diabetes in the offspring, with a vitamin D receptor–genotype interaction [46]. In contrast, average maternal vitamin D plasma concentration during pregnancy was not associated with an increased risk of type 1 diabetes in cohorts from Norway and Denmark [19]. Further data from prospective studies commencing in pregnancy, such as the Environmental Determinants of Islet Autoimmunity (ENDIA) study (anzctr.org.au registration no. ACTRN12613000794707) [21], will provide more comprehensive data on the role of perinatal vitamin D supplementation in IA and type 1 diabetes.

Cow’s milk protein

DAISY demonstrated that greater intake of cow’s milk protein was associated with increased IA risk in children with low/moderate risk HLA-DR genotypes (adjusted HR 1.41 [95% CI 1.08, 1.84]), but not in children with high-risk HLA-DR genotypes [10]. The Trial to Reduce Insulin-Dependent Diabetes Mellitus in the Genetically at Risk (TRIGR), an RCT that only recruited children with high-risk HLA genotypes, demonstrated that cow’s milk ingestion during infancy did not reduce the incidence of IA [34] or type 1 diabetes [35], as compared with hydrolysed formula. In the Diabetes Prediction and Prevention (DIPP) study, exposure to cow’s milk protein formula before 3 months of age was not associated with IA or type 1 diabetes, but the interaction between enterovirus infection and cow’s milk exposure before the age of 3 months was associated with IA [15]. If cow’s milk protein has a role in the pathogenesis of type 1 diabetes, the relationship may result from an interplay between genetic predisposition to aberrant mucosal immunity to dietary and other proteins [11], along with increased intestinal permeability (‘leaky gut’) and exposure to intestinal microbiota [47].

Breastfeeding

A meta-analysis of 43 studies, including 9874 cases of type 1 diabetes, demonstrated a weak protective effect of breastfeeding on type 1 diabetes risk: for exclusive breastfeeding ≥2 weeks vs <2 weeks, the OR was 0.75 (95% CI 0.64, 0.88) and for exclusive breastfeeding ≥3 months vs <3 months, the OR was 0.87 (95% CI 0.75, 1.00). The authors concluded that their findings were difficult to interpret due to marked heterogeneity and possible biases, particularly recall bias [48]. For ethical reasons, it will not be possible to address the question of exclusive breastfeeding in an RCT; however breastfeeding should be encouraged in children at risk of type 1 diabetes, as for the general population, due to other established benefits [49].

Gluten

Both early and delayed introduction of gluten have been implicated in the risk of IA and type 1 diabetes [6, 28]. In TEDDY (which included ~700 children with IA), early introduction of gluten before the age of 4 months reduced IA risk (adjusted HR 0.67 [95% CI 0.54, 0.98]), while late introduction of gluten (after 9 months of age) was not associated with increased IA risk [28]. Similarly, the BABYDIET study showed that delaying gluten exposure until 12 months of age in at-risk children was not associated with reduced IA risk [7]. The modest HRs in TEDDY and small sample size in BABYDIET suggest that further investigation into the effects of timing of gluten introduction on IA/type 1 diabetes risk is warranted in other cohorts, as well as studies into the interaction of gluten intake with other IA/type 1 diabetes risk variables, such as breastfeeding and infection.

Specific foods and micro- and macro-nutrients

Additional dietary factors that have been variably associated with risk of IA and type 1 diabetes, such as root vegetables [16, 17] or berries and fruit [16], were not shown to have significant associations with these conditions in the TEDDY dietary analyses [28]. Only the introduction of egg before 9 months of age was associated with a reduced IA risk (adjusted HR 0.8 [95% CI 0.72, 0.99]; p = 0.035). While the number of individuals in TEDDY [28] is much larger than the earlier studies in this area (highlighting the potential for small-study effects [50]), the discrepant findings between these studies may be related to chance findings in any of the studies, particularly given the large number of dietary variables measured and the lack of adjustment for multiple
comparisons [16, 28]. Other possible explanations include geographic differences in study populations, methodological differences between studies, changes in dietary practices over time and interaction between diet and other associated biological systems, such as the metabolome and gut microbiome.

**Overweight/obesity**

The rising incidence of type 1 diabetes largely parallels population rates of overweight/obesity over time. Using combined data from two large birth cohort studies in Norway and Denmark, type 1 diabetes risk in the offspring was modestly associated with maternal pre-pregnancy obesity (adjusted HR 1.41) and paternal obesity (adjusted HR 1.51) [20]. A meta-analysis of 29 predominantly European studies demonstrated that birthweight >3.5 kg was associated with an increased risk of type 1 diabetes, after adjustment for potential confounders [51]. In the Australian BABYDIAB study, weight and BMI z scores during infancy were continuous predictors of IA risk, with a weight z score >0 conferring a more than twofold risk of IA (HR 2.6) [4]. The effect of exposure to being overweight pre-conception, in utero and postnatally may be mediated by accompanying insulin resistance [52, 53].

**Infections**

An infectious aetiology for type 1 diabetes has been postulated for almost a century [54]. Multiple prospective cohort studies have investigated the role of early-life infections in type 1 diabetes, using prospective clinical data, as well as specific microbial and virological testing [23–26, 29, 55]. Recently, our understanding of the role of infections has been enhanced by the use of ‘omics’ technologies to determine the relationship between the microbiome (including bacteria, bacteriophages and viruses), transcriptome and proteome in the development of IA/type 1 diabetes.

**Maternal virus infection**

Pregnancy represents a complex immunological state in which bias towards helper T lymphocyte-associated cytokines (Th2 bias) diminishes cell-mediated immunity and increases vulnerability to intracellular infections, including viruses [56, 57]. Acute perinatal infections, such as rubella and infection with cytomegalovirus, can cause significant perinatal morbidity, with resulting clinical syndromes including diabetes; although most cases of congenital rubella syndrome have an atypical form of diabetes without evidence of IA [58]. In our systematic review and meta-analysis of 2039 mothers and 953 offspring, we found a significant association between virus infection during pregnancy and childhood type 1 diabetes (OR 2.2) but not IA [55]. The TEDDY study showed that retrospectively reported maternal respiratory infections during pregnancy were associated with a reduced risk of developing insulin autoantibodies in children with the CTLA4 G allele polymorphism (HR 0.64 [95% CI 0.45, 0.91]), suggesting a protective role for this allele [30]. Our systematic review highlighted the need for data from prospective cohorts commencing in pregnancy, with frequently obtained clinical samples to document infection, along with pathogenesis studies, to establish an aetiological link between in utero infections and type 1 diabetes in the offspring. The ongoing ENDIA study will address these questions, with a strong focus on ‘omics’ [21].

**Childhood infections**

Infections in early life increase the risk of IA and type 1 diabetes in at-risk and unselected population cohorts [23, 24, 26, 29]. In TEDDY, prospectively collected parent-reported history of recent respiratory infections in children before 4 years of age was associated with a modestly increased (5.6%) IA risk. The incidence of both IA and respiratory infections peaked at 6–9 months of age (Fig. 1) and a higher rate of respiratory infections was observed in the 9 months prior to seroconversion to IA [29]. In a population-based cohort from Germany, type 1 diabetes risk was significantly increased in children who experienced respiratory tract infections before 6 months of age [23]. Similarly, in an Australian birth cohort study of healthy children, type 1 diabetes risk was increased in association with an early upper respiratory tract infection before the age of 5 weeks (adjusted OR 2.7) or ear infection by the age of 12 weeks (adjusted OR 3.4) [26]. Since these studies were based on parental
reporting of infection, causative agents were not identified. In contrast, a large population-based study of >1 million cases and controls from Taiwan demonstrated that enterovirus infection before 18 years of age was associated with an increased risk of subsequent type 1 diabetes (adjusted HR 1.5) [25].

Such studies cannot determine whether the associations between early infections and IA or type 1 diabetes reflect increased exposure to viruses or an underlying susceptibility to virus infection due to a dysregulated immune response [59]. This may have a genetic basis, as antiviral response genes, such as IFIH1 CD69, and IL2RA, are among the loci linked to type 1 diabetes susceptibility [60]. In support of this hypothesis, a type I interferon-inducible transcriptional signature was increased in the peripheral blood of children from the BABYDIET study before seroconversion to IA and correlated with recent self-reported respiratory infections [8]. Also, it is possible that treatment with antibiotics in early life [61], which alter the gut microbiome, may explain the relationship between infection and IA/type 1 diabetes; although, antibiotic treatment likely serves as a confounder. Infections may contribute to IA and type 1 diabetes in young children through activation of interferon signalling pathways and pattern recognition receptors following bacterial, viral and, possibly, other infections [59]. This may result in beta cell inflammation, activation of autoreactive and bystander CD8+ T cells and progressive autoimmune-mediated beta cell destruction. The detection of IFN-α and increased HLA class I expression in beta cells of individuals with type 1 diabetes is consistent with this model [62].

**Viruses**

Numerous clinical, epidemiological and experimental studies have reported associations between IA or type 1 diabetes and viruses, particularly enteroviruses [63], but also herpesviruses (cytomegalovirus, Epstein–Barr virus) [64, 65], rotavirus [5], rubivirus (rubella) [66], rubulavirus (mumps) [66] and parechoviruses [67] during infancy and childhood. Potential mechanisms by which viruses interact with genetic predisposition to induce an autoimmune assault on pancreatic beta cells have been comprehensively reviewed elsewhere [59, 63].

**Enteroviruses**

In our systematic review and meta-analysis of 26 studies involving over 4400 participants, we found a significant association between enterovirus infection, IA and type 1 diabetes; for newly diagnosed type 1 diabetes, the OR was 12.7 (95% CI 6.4, 25.0) [68]. There was significant heterogeneity by geographical region, but we could not stratify the analysis by enterovirus genotype because the majority of studies did not provide this information.

DIPP demonstrated a temporal association between the appearance of the first diabetes-associated autoantibodies and enterovirus infection, based on both serology and PCR for enterovirus genes [69–71]. In particular, neutralising antibodies to coxsackievirus group B (CVB)1 increased the risk of IA, which was attenuated by the presence of maternal antibodies to the virus. In contrast, antibodies to CVB3 and CVB6 were associated with lower IA risk, which may be due to immunological cross protection against CVB1. These associations were also present in children who progressed to type 1 diabetes [70].

**Rotaviruses**

Recent reports from Australia and the USA have demonstrated a reduced incidence of type 1 diabetes in young children following the introduction of rotavirus vaccination [72, 73].

**Virome**

Although strong evidence supports a predominant role for enteroviruses in the pathogenesis of type 1 diabetes [68], the use of targeted detection methods (PCR, serology, etc.) raises significant concern for investigation bias towards enteroviruses. With the advent of high-throughput sequencing technologies, it is now possible to screen for all viruses (the ‘virome’) simultaneously, eliminating such bias. However, early application of virome sequencing in TEDDY [31] and DIPP [18] proved difficult due to the limited sensitivity caused by the overwhelming background of non-viral nucleic acid in clinical specimens [74].

To significantly improve the sensitivity of virome detection by up to 10,000-fold [74], we and others applied a novel viral enrichment strategy (VirCapSeq- VERT) to characterise the virome of children with IA in the Viruses In the Genetically At Risk (VIGR) study [36] and BABYDIET [9]. The sensitivity and specificity for virus detection using this method is on par with targeted PCR, unlike other virome-sequencing methods that use physical enrichment, which are less sensitive [36, 74]. In our analysis of the gut and plasma virome of 45 individuals with IA and 48 matched control participants, we detected 28 genera of viruses and 62% of children were positive for ≥1 vertebrate-infecting virus [36]. This represents a more than threefold higher positivity than previously reported in the DIPP virome analysis [18]. We identified 129 viruses as differentially abundant in the gut of individuals with IA vs control participants, including five enterovirus A genotypes, which were significantly more abundant in those with IA. While the sample size is small, these findings further support the contribution of enteroviruses to the development of IA and corroborate the proposal that viral load may influence IA risk [36].

In our analysis of the longitudinal gut virome of 61 pregnant women (35 with type 1 diabetes and 26 without) in ENDIA, using VirCapSeq- VERT [22], 63% of samples tested positive for at least one virus and 29 genera of eukaryotic viruses were detected (Fig. 2). Moreover, there was more than a twofold difference in the abundance of 77 viruses between the two maternal groups, including eight enterovirus B types, which were present at a higher abundance in women with type
1 diabetes. These findings provide novel insight into the diversity and dynamics of the gut virome during pregnancy and demonstrate a distinct profile of viruses during pregnancy in women with type 1 diabetes. It will be important to determine whether these differences remain significant in the full ENDIA cohort, and whether they translate to altered risk for IA and type 1 diabetes in the offspring. These findings may be informative for prevention strategies, such as enterovirus vaccination [63].

**Phageome** Gut bacteriophages are relatively unexplored in the pathogenesis of type 1 diabetes. Compared with eukaryotic viruses, they are far more ubiquitous, constituting the vast majority of viral sequences in the gut [18]. Analysis of the gut phageome in children with IA in the DIABIMMUNE study revealed significantly lower richness and Shannon diversity of bacteriophages in individuals with IA vs matched control participants [12]. Furthermore, a subset of bacteriophage sequences was directly or inversely associated with
seroconversion to IA and highly predictive for discriminating between cases and controls. Interestingly, these differences in bacteriophage populations before seroconversion were observed despite the lack of distinguishing bacterial signatures from the same samples [75]. This suggests that changes in the phageome prior to seroconversion may contribute to changes in bacterial diversity after seroconversion.

**Bacterial microbiome**

Dysbiosis of the gut microbiota has been associated with IA and type 1 diabetes, yet the functional consequences to the host of this dysbiosis are still being unravelled. The composition of the gut microbiome is strongly affected by the maternal environment, diet (particularly breastfeeding in early life) and use of antibiotics and probiotics. The microbiome interacts closely with the host immune system, making it a strong contender for providing a mechanistic link between multiple environmental modifiers and IA.

**Maternal microbiome**

A landmark study of the gut microbiome in healthy human pregnancy showed marked changes from the first to the third trimester, to a more proinflammatory microbiome [76]. Such changes could be considered adaptive to support the growing fetus and weight gain of pregnancy; however, these changes were not replicated in subsequent studies of pregnancy, with reports of a stable gestational microbiome across different body sites [77]. Maternal and, potentially, paternal factors influence the infant’s microbiome after rapid changes during the first year of life. Maternal weight, pregnancy weight gain and complications of pregnancy, such as diabetes and pre-eclampsia, all influence the composition and functional capacity of the gut microbiome, as followed through preschool years [32]. Antibiotic use in pregnancy can affect the maternal and infant microbiome, but overall exposure, number of courses or antiobiotic type (narrow vs broad) in utero were not associated with type 1 diabetes risk [78].

**Infant microbiome**

The gut microbiome has a critical role in the early development of immune regulation in the infant [79]. The early microbiome, and its perturbations in association with IA and type 1 diabetes, is the subject of investigation within multiple cohorts. Whilst most studies used 16S ribosomal RNA (rRNA) sequencing to assess community composition, modern metagenomic methods more accurately identify functional and strain-specific differences in the microbiome [13] and other approaches, such as faecal metaproteomics, can identify host–microbiota interactions [80]. Microbial composition during early life is influenced by mode of birth, breastfeeding, introduction of formula milk and solid foods, antibiotic and probiotic use, and bacterial and viral infections. A decrease in diversity and stability of case microbiomes over time, as well as a reduction in short-chain fatty acid (SCFA) production and gut integrity-associated bacterial genes, were associated with IA and type 1 diabetes [33, 75]. In DIABIMMUNE, alpha diversity decreased after seroconversion in individuals who progressed from IA to type 1 diabetes, with the onset of type 1 diabetes being preceded by inflammation-associated microbes and functional pathways [75].

The recent longitudinal analysis of stool microbiomes in TEDDY (mainly white non-Hispanic children), using 16S rRNA and metagenomic sequencing, is the largest study to examine the infant microbiome [32, 33]. Three phases were described: (1) a developmental phase (3–14 months); (2) a transitional phase (15–30 months); and (3) a stable phase (31–46 months). Partial or exclusive breastfeeding was the most significant factor associated with microbial composition, resulting in increased *Bifidobacterium* species. Weaning led to faster maturation of the gut microbiome, driven by the breast milk cessation, heralding the appearance of bacteria belonging to the phylum Firmicutes. Other factors that significantly associated, but to a lesser degree, with infant microbiota composition were geographical location, antibiotic use, probiotic use, household siblings, furry pets and vitamin D supplementation [32].

Further, breast milk feeding between 3 and 14 months was solely associated with bacterial metabolic potential. Pathways involving metabolic functions associated with the adult microbiome were increased in non-breastfed infants up to 14 months of age [32]. A reduction in gene families involved in the degradation of human milk oligosaccharides to SCFA and an increase in mucin-degrading genes after weaning indicated a strain-shift, whereby new microbiota, including *Enterococcus* species, were increasingly responsible for these functions [78].

Relatively modest alterations in the microbial composition were associated with IA or type 1 diabetes in both the amplicon sequencing [32] and metagenomic sequencing [33] arms of TEDDY. Control participants had higher levels of *Lactococcus* and *Streptococcus*, both of which are common in dairy products. Gut microbiomes of children in the TEDDY study who did or did not develop IA or type 1 diabetes, therefore, did not reveal clear taxonomic differences and there was considerable geographical heterogeneity across the study sites. However, importantly, on metagenomic analysis, the microbiomes of progressors to IA or type 1 diabetes contained significantly higher numbers of genes involved in pathways involved in fermentation and production of SCFA byproducts. SCFA bacterial products regulate host metabolism and immunity. Butyrate, in particular, maintains gut epithelial integrity,
supports differentiation of regulatory T cells and promotes anti-inflammatory responses, and both butyrate and acetate have been shown to prevent diabetes in the NOD mouse. The TEDDY study findings, therefore, extend experimental data in the NOD mouse [81] and smaller human studies [82] that support a protective effect of SCFA on the development of type 1 diabetes. Interestingly, there was geographical variability as to which SCFA pathways (acetate, butyrate or propionate) were altered in children with IA/type 1 diabetes in the TEDDY study, and the taxa with which these were associated. This suggests that changes in the overall molecular functions of the microbiota may be more uniformly associated with type 1 diabetes and IA, rather than specific taxa. A recent study utilising metaproteomic analysis of stool samples sheds light on the functional impact of altered microbiota on the host; the abundance of proteins from the mucous barrier and epithelial adhesion molecules were positively associated with taxa that were abundant in control participants compared with individuals with IA or type 1 diabetes [80]. These findings further support gut epithelial integrity as a key characteristic of the microbiome in healthy individuals.

Despite TEDDY’s large sample size, most associations were modest in effect size, which may be attributed to differing type 1 diabetes endotypes, high geographical heterogeneity and a lack of samples from early life (prior to 3 months of age) [33]. The impact of the microbiome and its products may be different with different genetic risk alleles and in different disease phenotypes. The role of other co-inhabitants, such as the mycobiome, is unknown, as is the impact of the dramatic changes in biological systems before and after birth. Studies investigating pregnancy to early-life cohorts, such as ENDIA [21] and DIPP Novum, will address these unknowns.

Crosstalk between early-life factors and pancreatic beta cells

While a diverse spectrum of early-life factors putatively influence the development of IA and type 1 diabetes, the mechanisms underlying their contribution to the autoimmune assault that specifically targets pancreatic beta cells are yet to be fully elucidated [59]. One potential route is the induction of the beta-cell in the endoplasmic reticulum (ER), where the synthesis/folding of proinsulin takes place [83]. During insulitis, proinflammatory cytokines released by infiltrating immune cells disrupt ER homeostasis, triggering the unfolded protein response. This adaptive phase is believed to play a critical role in the development of IA by promoting the generation of neoantigens and activating inflammatory responses [84]. Furthermore, ER stress induces translational errors, giving rise to defective ribosomal products of insulin that are highly immunogenic and targeted by T cell autoreactivity in type 1 diabetes [85].

Research gaps and future directions

The discrepant findings from studies investigating the multitude of putative aetiological determinants of IA and type 1 diabetes, with largely modest associations observed, highlight the importance of adequately powered studies, consistent definitions of exposure and outcome measures, multifactorial modelling of pathogenesis and meta-analyses, which can also explore publication bias and study heterogeneity. Findings from small or pilot studies may be chance associations, particularly when not replicated in larger cohorts or trials, as may be the case for TRIGR [35]. However, large studies and ‘big data’ are not immune from bias or chance findings [86]. Key to unravelling the heterogeneous nature of type 1 diabetes is the exploration of genetic susceptibility and validation of GRSs in diverse populations. Studies need to take account of the interplay between development of islet autoimmunity and/or type 1 diabetes and environmental factors, stratified by genetic, ethnic and global variation, particularly in early life. While incidence rates for type 1 diabetes are highest in Scandinavia/Northern Europe [87], the caseload of type 1 diabetes globally includes regions that have not been comprehensively studied, such as the Middle East, Africa, Asia and South America. For interventions aimed at primary, secondary and tertiary prevention of type 1 diabetes to be successful, understanding the heterogeneity of type 1 diabetes is essential. In the era of ‘omics’, the future potential to elucidate the interaction between multiple biological systems in IA/type 1 diabetes is immense.

Funding MEC is supported by an Australian National Health and Medical Research Council (NHMRC) Practitioner fellowship (APP1045777) and a JDRF International grant (3-SRA-2017-417-A-N). SRI is supported by an Australian Government Research Training Program Scholarship. EEH-W is supported by grants from JDRF (3-SRA-2019-730-S-B, 2-SRA-2019-703-M-B) and a Children’s Hospital Foundation grant (WIS0202018). The ENDIA Study is supported by JDRF Australia, the recipient of the Australian Research Council Special Research Initiative in Type 1 Juvenile Diabetes, The Leona M. and Harry B. Helmsley Charitable Trust, JDRF International and the JDRF/NHMRC Centre of Research Excellence for the Protection of Pancreatic Beta Cells (APP1078106; JJC, MEC, MP, WDR).

Duality of interests The authors declare that there is no duality of interest associated with this manuscript.

Contribution statement All authors contributed to the conception and design, interpretation and critical appraisal of data, drafting of the review and revising it critically for important intellectual content. All authors approved of the version to be published.

References


Springer


Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
Higher frequency of vertebrate-infecting viruses in the gut of infants born to mothers with type 1 diabetes

Ki Wook Kim | Digby W. Allen | Thomas Briese | Jennifer J. Couper
Simon C. Barry | Peter G. Colman | Andrew M. Cotterill | Elizabeth A. Davis
Lynne C. Giles | Leonard C. Harrison | Mark Harris | Aveni Haynes
Jessica L. Horton | Sonia R. Isaacs | Komal Jain | Walter I. Lipkin
Kelly McGorm | Grant Morahan | Claire Morbey | Ignatius C. N. Pang
Anthony T. Papenfuss | Megan A. S. Penno | Richard O. Sinnott
Georgia Soldatos | Rebecca L. Thomson | Peter Vuillermin
John M. Wentworth | Marc R. Wilkins | William D. Rawlinson
Maria E. Craig

1School of Women’s and Children’s Health, University of New South Wales, Sydney, New South Wales, Australia
2Center for Infection and Immunity and Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, New York
3Robinson Research Institute and Adelaide Medical School, University of Adelaide, Adelaide, South Australia, Australia
4Department of Diabetes and Endocrinology, The Royal Melbourne Hospital Victoria, Melbourne, Victoria, Australia
5Department of Endocrinology, Queensland Children’s Hospital, South Brisbane, Queensland, Australia
6Telethon Kids Institute, The University of Western Australia, Perth, Western Australia, Australia
7School of Public Health, University of Adelaide, Adelaide, South Australia, Australia
8Walter and Eliza Hall Institute and Royal Melbourne Hospital, Melbourne, Victoria, Australia
9Centre for Diabetes Research, Harry Perkins Institute of Medical Research, Perth, Western Australia, Australia
10Hunter Diabetes Centre, Newcastle, New South Wales, Australia
11School of Biotechnology and Biomolecular Science, University of New South Wales, Sydney, New South Wales, Australia
12Department of Computing and Information Systems, University of Melbourne, Melbourne, Victoria, Australia
13Monash Centre for Health Research and Implementation, School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia
14School of Medicine, Deakin University, Geelong, Victoria, Australia
15Serology and Virology Division, SEALS Microbiology, Prince of Wales Hospital, Sydney, New South Wales, Australia
16Institute of Endocrinology and Diabetes, The Children’s Hospital at Westmead, Sydney, New South Wales, Australia

Correspondence
Prof Maria E. Craig, Institute of Endocrinology and Diabetes, The Children’s Hospital at Westmead, Locked Bag 4001, Westmead NSW 2145, Australia.
Email: m.craig@unsw.edu.au

Funding information
Juvenile Diabetes Research Foundation Australia, Grant/Award Number: 4-SRA-

Abstract
Background: Microbial exposures in utero and early life shape the infant microbiome, which can profoundly impact on health. Compared to the bacterial microbiome, very little is known about the virome. We set out to characterize longitudinal changes in the gut virome of healthy infants born to mothers with or without type 1 diabetes using comprehensive virome capture sequencing.
Methods: Healthy infants were selected from Environmental Determinants of Islet Autoimmunity (ENDIA), a prospective cohort of Australian children with a first-degree relative with type 1 diabetes, followed from pregnancy. Fecal specimens were collected three-monthly in the first year of life.

Results: Among 25 infants (44% born to mothers with type 1 diabetes) at least one virus was detected in 65% (65/100) of samples and 96% (24/25) of infants during the first year of life. In total, 26 genera of viruses were identified and >150 viruses were differentially abundant between the gut of infants with a mother with type 1 diabetes vs without. Positivity for any virus was associated with maternal type 1 diabetes and older infant age. Enterovirus was associated with older infant age and maternal smoking.

Conclusions: We demonstrate a distinct gut virome profile in infants of mothers with type 1 diabetes, which may influence health outcomes later in life. Higher prevalence and greater number of viruses observed compared to previous studies suggests significant underrepresentation in existing virome datasets, arising most likely from less sensitive techniques used in data acquisition.

KEYWORDS
enterovirus, gut, longitudinal, type 1 diabetes, virome

1 | INTRODUCTION

Microbial exposures in utero and early life shape the infant microbiome, which can profoundly impact on development, health, and immune maturation.1,2 Although the human microbiome encompasses a diverse population of bacteria, archaea, protists, fungi, viruses, and bacteriophages, current understanding stems mostly from characterization of the gut bacterial population using targeted sequencing of the 16S ribosomal RNA gene.3-7 In contrast to bacteria, viruses lack a universally conserved genetic region that can be targeted for amplification or enrichment. Furthermore, viruses are present at a significantly lower abundance in the gut compared to bacteria or their phages. Thus, it is challenging to characterize the complete population of viruses (“virome”) using conventional metagenomic sequencing.

All published infant virome studies to date have performed metagenomic sequencing without a highly effective method for virus enrichment.8-12 Most used a combination of physical enrichment techniques such as filtration, centrifugation and nuclease treatment that only provide modest increases in viral reads.13 Although these approaches provided sufficient sensitivity to explore the diversity and frequency of the highly abundant bacteriophage population, it is unclear whether this was also the case for non-phage viruses which were seldom detected.

Here we hypothesized that viruses are significantly underrepresented in existing infant virome datasets, and that infants born to a mother with type 1 diabetes have a distinct gut virome profile compared to those from a mother without diabetes. We tested this and further elucidated the composition, richness and the dynamics of viruses in the infant gut through comprehensive virome capture sequencing (VirCapSeq-VERT) of longitudinal fecal samples collected from 25 healthy infants during the first year of life. VirCapSeq-VERT enables effective enrichment of sequences corresponding to all viruses known to infect vertebrates, increasing viral read recovery up to 10,000-fold compared to conventional metagenomic sequencing.14

2 | METHODS

2.1 | Study subjects and sample selection

The case-control study population was nested within Environmental Determinants of Islet Autoimmunity (ENDIA), a longitudinal early life prospective cohort of children with at least one first-degree relative diagnosed with type 1 diabetes followed from pregnancy. We selected 25 infants who had fecal specimens collected at birth and 3, 6, 9, and either 12 or 15 months of age (Tables 1 and 2). In total, 100 longitudinal fecal samples stored at −80 °C were examined by VirCapSeq-VERT. None of the infants have developed type 1 diabetes.

The study was reviewed and approved (13 July 2016) by the Human Research Ethics Committee at the Women’s and Children’s Health Network under the National Mutual Acceptance Scheme (HREC/16/WCHN/66) and at all participating study sites in Australia. All participants provided written informed consent and were free to withdraw from the study at any time. Families were excluded if the mother could not comprehend her participation in the study and therefore was unable to provide informed consent.
TABLE 1  Characteristics of infants stratified by maternal type 1 diabetes status

<table>
<thead>
<tr>
<th></th>
<th>Infants of mothers with type 1 diabetes (N = 11)</th>
<th>Infants of mothers without type 1 diabetes (N = 14)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male, n (%)</td>
<td>5 (46)</td>
<td>5 (36)</td>
<td>.7</td>
</tr>
<tr>
<td>Season of birth n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summer</td>
<td>2 (18)</td>
<td>1 (7)</td>
<td>.6</td>
</tr>
<tr>
<td>Autumn</td>
<td>2 (18)</td>
<td>4 (29)</td>
<td>.7</td>
</tr>
<tr>
<td>Winter</td>
<td>3 (28)</td>
<td>4 (29)</td>
<td>.0</td>
</tr>
<tr>
<td>Spring</td>
<td>4 (36)</td>
<td>5 (33)</td>
<td>1.0</td>
</tr>
<tr>
<td>Pet ownership, n (%)</td>
<td>5 (45)</td>
<td>10 (71)</td>
<td>.2</td>
</tr>
<tr>
<td>Siblings, n (%)</td>
<td>4 (36)</td>
<td>8 (57)</td>
<td>.4</td>
</tr>
<tr>
<td>Number of siblings, mean ± SD</td>
<td>.0.4 ± 0.5</td>
<td>.0.7 ± 0.7</td>
<td>.2</td>
</tr>
<tr>
<td>Obstetric history of infant</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Birth weight (kg), mean ± SD</td>
<td>3.4 ± 0.6</td>
<td>3.5 ± 0.5</td>
<td>.8</td>
</tr>
<tr>
<td>Weight for age z-score, mean ± SD</td>
<td>1.5 ± 2.3</td>
<td>0.2 ± 1.1</td>
<td>.07</td>
</tr>
<tr>
<td>Gestation (wk), mean ± SD</td>
<td>37.2 ± 1.3</td>
<td>39.2 ± 1.8</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Vaginal delivery, n (%)</td>
<td>7 (64)</td>
<td>8 (57)</td>
<td>1.0</td>
</tr>
<tr>
<td>Maternal demographics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age at conception (y), mean ± SD</td>
<td>30.8 ± 3.9</td>
<td>33.5 ± 4.4</td>
<td>.02</td>
</tr>
<tr>
<td>High SES, n (%)</td>
<td>7 (64)</td>
<td>7 (50)</td>
<td>.7</td>
</tr>
<tr>
<td>Pre-pregnancy BMI, mean ± SD</td>
<td>26.2 ± 4.0</td>
<td>25.3 ± 4.8</td>
<td>.3</td>
</tr>
<tr>
<td>Tertiary education, n (%)</td>
<td>10 (91)</td>
<td>12 (86)</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Abbreviations: BMI, body mass index; SES, socioeconomic status.

2.2 Virome sequencing

Protocols for nucleic acid extraction, cDNA synthesis, sequence-independent amplification and VirCapSeq-VERT have been described previously. Briefly, total nucleic acid was extracted using the MagMAX Total Nucleic Acid Isolation Kit (Thermo Fisher Scientific, Waltham, Massachusetts) from 30% (w/v) fecal suspensions prepared in 1× PBS. To generate 1 μg double-stranded DNA for library synthesis, total nucleic acid was subjected to cDNA synthesis and sequence-independent pre-amplification using the Transplex Complete Whole Transcriptome Amplification Kit (Sigma-Aldrich, WTA1, St. Louis, Missouri). Following amplification, purified products were used to prepare Illumina sequencing libraries using the KAPA Hyperplus kit (KAPA Biosystems, Wilmington, Massachusetts) with single-index adapters, compatible for Sequence Capture Enrichment (Roche, Basel, Switzerland). VirCapSeq-VERT was performed at Columbia University, New York. Capture was performed according to the Nimblegen SeqCap protocol (Roche, Basel, Switzerland) as outlined previously. Post-capture, virus-enriched libraries were purified and amplified before sequencing. To ensure sufficient depth of coverage (about 10 million raw sequence reads/sample), uniquely barcoded samples were pooled at a maximum of 20 libraries per pool for sequence capture and each pool was sequenced single-end on an individual lane of HiSeq2500 (Illumina, San Diego, California).

2.3 Metagenomic sequence analysis

Sequence trimming, host sequence filtration, generation of contiguous read assemblies (contigs) and taxonomic classification of reads were performed as previously described. In brief, de-multiplexed and quality-trimmed sequence reads were aligned against host reference databases from GenBank (NCBI) using the Bowtie2 mapping algorithm (version 2.1.0) to remove the host background. Filtered reads were de novo assembled using either SOAPdenovo2, MEGAHIT or MIRA assemblers, then contigs and unique singletons were subjected to homology search at the nucleotide level using megablast. Sequences that exhibited poor or no homology at the nucleotide level were screened further using BLASTX against the viral GenBank protein database. For reference-based alignments, visualization of depth and spread of coverage for individual viruses, both Integrated Genomics Viewer21 and Geneious (version 9.0.5) were used.

2.4 Statistical analysis

Continuous demographic variables are reported as a mean ± SD (SD) for symmetrically distributed variables and median [IQR] for skewed variables. Categorical variables are summarized as number (%). Participant characteristics, including demographic variables, lifestyle factors, and comorbidities were stratified based on maternal type 1 diabetes and compared using independent t tests and Fisher’s exact tests for continuous and categorical variables, respectively. The socioeconomic index for areas (SEIFA) percentile for the postal area in which each participant resided was used as an indicator of socioeconomic status (SES). High SES was defined as >75th percentile and low SES was defined as <50th percentile.

Consistent with previous virome analyses, the virus positivity threshold was set at ≥100 viral reads matched by BLAST at the species level after a conservative 1% correction for potential sequence bleed-through. The estimated rate of bleed-through on the Illumina platform is ~0.3% of total reads when using a single index adapters. Associations between virus positivity and explanatory variables were examined using univariate and multivariable generalized estimating equations (GEE). Logistic regression models for the binary outcome of virus positivity at genus level were fit, with GEE used to account for the correlation among serial observations from the same infant. The major explanatory variable considered was maternal type 1 diabetes. Other explanatory variables investigated were maternal age at conception, parity, pet ownership, SES, pre-pregnancy body mass index (BMI),
1 diabetes was examined using the edgeR package (version 3.14.0) in R (version 3.3.0). A matrix of read counts was generated encompassing all samples and detected viruses. Each matrix entry had a count of one per sample following host and primer sequence removal. In total, 26 genera of viruses were detected and 65% (65/100) of samples were positive for nucleic acid corresponding to at least one virus. This equated to 96% (24/25) of infants being virus-positive in the first year of life. Noroviruses (28% of all samples positive), enteroviruses (26%), parechoviruses (14%), anelloviruses (11%), and bocaparvoviruses (7%) were the most frequently detected, and multiple viruses were detected in 31% of samples (Figure 1). In 60% (15/25) of infants, viruses of the same genus were detected at multiple timepoints across consecutive visits or 6 months apart (Figure 2). Viral richness was significantly lower in earliest-in-life specimens compared to the latest timepoint ($P = .0034$).

Of the 25 infants examined, 11 were born to mothers with type 1 diabetes; characteristics stratified by maternal diabetes status are shown in Table 1. Mothers with type 1 diabetes were younger at conception and had shorter gestation. Total virus positivity was higher in infants of mothers with type 1 diabetes compared to those without (75% vs 59%). In univariate analysis, positivity for any virus was associated with older infant age (OR 1.2, 95% CI 1.1 to 1.3, $P = .002$) while the association with maternal type 1 diabetes did not reach statistical significance (OR 2.1, 95% CI 0.9 to 4.7, $P = .07$). Enterovirus positivity was associated with older age (OR 1.1, 95% CI 1.0 to 1.3, $P = .032$) and maternal smoking (OR 2.8, 95% CI 1.8 to 4.4, $P < .0001$), norovirus with low SES (OR 4.0, 95% CI 1.2 to 13.0, $P = .02$) and anellovirus with greater number of siblings (OR 2.4, 95% CI 1.1 to 5.3, $P = .03$). Parechovirus was associated with pet ownership (OR 5.5, 95% CI 1.3 to 24.0, $P = .02$) and older age (OR 1.2, 95% CI 1.0 to 1.3, $P = .01$). In multivariable GEE analysis, positivity for any virus was associated with maternal type 1 diabetes (OR 9.2, 95% CI 2.2 to 39.0, $P = .003$) after adjustment for infant age. The interaction between infant age and maternal type 1 diabetes was significant ($P = .005$). No other examined variables including breastfeeding, mode of delivery and antibiotics intake were significantly associated with virus positivity.

Differential abundance analysis identified 165 viruses with a $\geq 2$-fold difference in viral read abundance between infants of mothers with vs without type 1 diabetes (Figure S1). However, some of these differences were based on viral reads detected from a single sample. When only viruses detected in $\geq 3$ samples were included in the analysis, 17 differentially abundant viruses

### TABLE 2  Characteristics of infants stratified by sample time point

<table>
<thead>
<tr>
<th>Maternal T1D (Y/N)</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
<th>Sample 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y (n)</td>
<td>N (n)</td>
<td>Y (n)</td>
<td>N (n)</td>
</tr>
<tr>
<td>Age (mo), mean ± SD</td>
<td>3.1 ± 3.1</td>
<td>1.8 ± 2.1</td>
<td>5.2 ± 1.3</td>
<td>4.9 ± 2.3</td>
</tr>
<tr>
<td>Breastfeeding, n (%)</td>
<td>10 (91)</td>
<td>14 (100)</td>
<td>8 (73)</td>
<td>14 (100)</td>
</tr>
<tr>
<td>Season, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summer</td>
<td>3 (27)</td>
<td>5 (36)</td>
<td>1 (8)</td>
<td>4 (29)</td>
</tr>
<tr>
<td>Autumn</td>
<td>2 (18)</td>
<td>1 (7)</td>
<td>4 (37)</td>
<td>6 (43)</td>
</tr>
<tr>
<td>Winter</td>
<td>4 (37)</td>
<td>3 (21)</td>
<td>2 (18)</td>
<td>2 (14)</td>
</tr>
<tr>
<td>Spring</td>
<td>2 (18)</td>
<td>5 (36)</td>
<td>4 (37)</td>
<td>2 (14)</td>
</tr>
<tr>
<td>Maternal smoking, n (%)</td>
<td>0 (0)</td>
<td>1 (7)</td>
<td>1 (9)</td>
<td>1 (7)</td>
</tr>
<tr>
<td>Pets, n (%)</td>
<td>4 (37)</td>
<td>10 (71)</td>
<td>5 (45)</td>
<td>10 (71)</td>
</tr>
</tbody>
</table>

3 | RESULTS

Virome sequencing of 100 fecal specimens from 25 infants generated approximately 1.8 billion raw reads. This translated to 14.8 ± 7.8 M reads...
remained. Among the 15 most differentially abundant viruses, human bocavirus, and rotavirus A were more abundant in the infants of mothers with type 1 diabetes (Table 3). Conversely, human parechoviruses, coxsackievirus A6, Rhinovirus C, and torque teno viruses were significantly less abundant. Some noroviruses were more abundant in infants of mothers with type 1 diabetes and others less abundant.

4 | DISCUSSION

We detected a diverse range of viruses in the infant gut during the first year of life, many of which were more prevalent than reported in previous virome studies. Although various factors can influence virus positivity during infancy, our observations likely reflect the enhanced sensitivity of VirCapSeq-VERT for detecting vertebrate-infecting viruses over conventional metagenomic sequencing. Indeed, our application of VirCapSeq-VERT to other pregnancy and childhood specimens also indicated higher virus positivity compared to previous virome studies.

Total virus positivity was positively associated with maternal type 1 diabetes and older infant age, with a significant interaction between these two variables. We speculate that infants of mothers with type 1 diabetes are more likely to harbor viruses in the gut compared to infants whose mothers do not have type 1 diabetes. In contrast, total virus positivity was not associated with the number or the presence of siblings. This suggests that a mother with type 1 diabetes may be a significant source of viruses for infants in the first year of life. The absence of matching timepoint maternal samples precluded investigation of whether mothers with type 1 diabetes are in general more likely to harbor viruses compared to those without diabetes. However, this is certainly possible given the trend to higher virus positivity in pregnant women with type 1 diabetes compared to pregnant women without diabetes and the higher rates of microbial infection observed in individuals with type 1 diabetes vs without.

Enteroviruses were detected in 26% of samples and in 76% (19/25) of infants during the first year of life. This represents more than double the prevalence of enteroviruses found in the gut virome analysis of 22 infants in the DIABIMMUNE study. In univariate analysis, enterovirus presence was associated with older infant age. This is consistent with the progressive increase of enterovirus prevalence in fecal samples in the first year of life. Although our sample size was small, enterovirus positivity was also associated with maternal smoking. As smoking impairs responsiveness to viral infections and cause immune dysfunction, the effects of maternal smoking on virus susceptibility in the offspring should be explored and validated in larger cohorts.

A large body of evidence supports the role of enteroviruses as prime environmental triggers of islet autoimmunity and type 1 diabetes. In our recent analysis of the pregnancy gut virome, the frequency of enteroviruses did not differ significantly between pregnant women with and without type 1 diabetes. Similarly, examination of the gut virome changes preceding the development of islet autoimmunity and type 1 diabetes in 11 case children by Zhao et al found no significant differences in enterovirus abundance or prevalence between cases and controls. In the present analysis, there was no difference in enterovirus positivity between infants born of a mother with type 1 diabetes vs without. However, the abundance of some enteroviruses differed significantly, including a higher abundance of...
Coxsackievirus A6 in infants of mothers without type 1 diabetes. Interestingly, we previously reported higher abundance enterovirus A species in the gut of children with islet autoimmunity. Therefore, infants positive for enterovirus A in this study will be followed closely for their progression to islet autoimmunity.

Norovirus was present in 28% of samples and associated with low SES in our cohort. As both SES and norovirus infection are commonly associated with a reduced level of hygiene, infants in low SES families may be exposed more frequently to noroviruses compared to infants from high SES households. In contrast, no other viruses were

**FIGURE 2** Longitudinal dynamics of the infant gut virome in the first year of life. Presence-absence heatmap of viruses detected in 25 infants (11 from mothers with type 1 diabetes) over four consecutive study visits (V1-4). Viruses are represented at the genus level.

**TABLE 3** Top 15 differentially abundant viruses (detected in ≥3 samples) between the gut of infants from mothers with vs without type 1 diabetes.
associated with SES. Parechovirus was associated with pet ownership but a zoonotic link between parechoviruses in animals and humans has yet to be demonstrated. Positivity of parechoviruses peaked between 6 and 12 months of age, which corresponds with the nadir of IgG as maternal antibodies decline.

Consistent with other studies, anelloviruses were frequently detected in infancy. In our analysis, anellovirus was associated with greater number of siblings, supportive of the notion that a key source of its exposure is through contact with other infants. This is consistent with the lack of evidence for the vertical transmission of anelloviruses. Although their pathogenicity in humans remains uncertain, the expansion of anelloviruses in the gut and blood correlates with diminished immune status in immunocompromised patients.

Interpretation of these results should take into consideration the following limitations. Firstly, as with all sequence-based virus detection, positivity for viral nucleic acid is a marker of, not proof of, infection. Viruses or viral nucleic acid may pass through the gut without causing a productive infection, as it is the case with plant viruses and other diet derived viruses. Second, we specifically focused on vertebrate-infecting viruses, excluding from our analysis other virus, bacteriophage and bacterial population coinhabiting the infant gut. This was deliberate to maximize the detection sensitivity for potentially human-infectious viruses. Although it would be desirable to obtain a wholistic view of the infant gut microbiome, recent evidence suggests that viruses have no influence on the bacterial microbiome or bacteriophage frequency. Third, all infants had a first degree relative with type 1 diabetes, who have a higher than baseline risk for development of islet autoimmunity and type 1 diabetes. Therefore, some of the associations identified in our univariate analysis may not reflect the general population. None of the infants have developed type 1 diabetes to date. Finally, our sample size precludes exploration of multiple variables that may act as confounders or effect modifiers in multivariable analyses. However, even if we select a more conservative p value of 0.01, the association with maternal diabetes and older age remains significant.

Strengths of our study include the first application of the highly sensitive VirCapSeq-VERT in a healthy infant cohort, the unbiased detection of all viruses simultaneously and the longitudinal nature of our cohort, which will enable validation of our findings using larger numbers through infancy and childhood. Furthermore, the availability of corresponding maternal pregnancy samples allows future investigation of whether any viruses detected in the present study were potentially vertically transmitted.

In conclusion, we provide a comprehensive and unbiased characterization of the gut virus population in infants during the first year of life and evaluate differences based on maternal type 1 diabetes. We identify older infant age, SES, sibling, pets, maternal smoking and maternal type 1 diabetes as factors influencing the gut virome in infants. Moreover, the higher prevalence and number of viruses observed compared to previous studies suggests an underrepresentation in previously reported virome datasets.

ACKNOWLEDGEMENTS

We thank Nishit P. Bhuvu (Columbia University, NY) for performing the virome capture hybridization and submitting samples for sequencing; Cheng Guo (Columbia University, NY) for assisting in the processing of metagenomic data and generation of Basic Local Alignment Search Tool (BLAST) result summaries. The authors acknowledge funding from the National Health and Medical Research Council (NHMRC) Practitioner fellowships (APP1045777 M.E.C. and APP1044684 W.D.R.), Juvenile Diabetes Research Foundation Australia (JDRF)/NHMRC Centre of Research Excellence for the Protection of Pancreatic Beta Cells (APP1078106), JDRF International (3-SRA-2017-417-A-N) and the National Institutes of Health (Center for Research in Diagnostics and Discovery) (U19 AI10976 T.B. and W.I.L.). The ENDIA Study (ANZCTR # ACTRN12613000794707) is supported by JDRF Australia, the recipient of the Australian Research Council Special Research Initiative in Type 1 Juvenile Diabetes, The Leona M. and Harry B. Helmsley Charitable Trust, JDRF International, and the NHMRC.

CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

AUTHOR CONTRIBUTIONS

M.E.C., W.D.R., T.B., W.I.L., and K.W.K. designed the study. K.W.K., S.R.I., and J.L.H. performed the experiments. I.C.N.P. performed the differential abundance analysis, K.J. and T.B. performed the de novo assembly of Illumina sequence reads and BLAST analysis. K.W.K., D.W.A., and M.E.C. performed the univariate and multivariable GEE analyses. All authors contributed to the interpretation of results. K.W.K., D.W.A., and M.E.C. wrote the manuscript, and all authors edited the manuscript.

ORCID

Jennifer J. Couper https://orcid.org/0000-0003-4448-8629
Aveni Haynes https://orcid.org/0000-0001-9954-5016
Rebecca L. Thomson https://orcid.org/0000-0002-7807-4144
Maria E. Craig https://orcid.org/0000-0001-6004-576X

REFERENCES


**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Kim KW, Allen DW, Briese T, et al. Higher frequency of vertebrate-infecting viruses in the gut of infants born to mothers with type 1 diabetes. *Pediatr Diabetes*. 2019;1–9. [https://doi.org/10.1111/pedi.12952](https://doi.org/10.1111/pedi.12952)
Respiratory viral co-infections among SARS-CoV-2 cases confirmed by virome capture sequencing

Ki Wook Kim1,2, Ira W. Deveson3,4, Chi Nam I. Pang6, Malinna Yeang2, Zin Naing2, Thiruni Adikari6,7, Jillian M. Hammond3, Igor Stevanovski2, Alicja G. Beukers8, Andrey Verich7, Simon Yin9, David McFarlane9, Marc R. Wilkins5, Sacha Stelzer-Braid2,6, Rowena A. Bull6,7, Maria E. Craig1,2,10, Sebastiaan J. van Hal8,11 & William D. Rawlinson1,2,5,6

 Accumulating evidence supports the high prevalence of co-infections among Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) patients, and their potential to worsen the clinical outcome of COVID-19. However, there are few data on Southern Hemisphere populations, and most studies to date have investigated a narrow spectrum of viruses using targeted qRT-PCR. Here we assessed respiratory viral co-infections among SARS-CoV-2 patients in Australia, through respiratory virome characterization. Nasopharyngeal swabs of 92 SARS-CoV-2-positive cases were sequenced using pan-viral hybrid-capture and the Twist Respiratory Virus Panel. In total, 8% of cases were co-infected, with rhinovirus (6%) or influenzavirus (2%). Twist capture also achieved near-complete sequencing (> 90% coverage, > tenfold depth) of the SARS-CoV-2 genome in 95% of specimens with Ct < 30. Our results highlight the importance of assessing all pathogens in symptomatic patients, and the dual-functionality of Twist hybrid-capture, for SARS-CoV-2 whole-genome sequencing without amplicon generation and the simultaneous identification of viral co-infections with ease.

Early description of the first 99 COVID-19 cases in Wuhan suggested that co-infections with other respiratory pathogens were rare1. However, more recent data from Northern California, USA demonstrate that rates of respiratory co-infections between SARS-CoV-2 and other respiratory viruses can reach up to 21%3. Furthermore, a higher prevalence of co-infection is reported among COVID-19 patients with more severe onset of disease3 and the deceased4, suggesting that co-infections can significantly worsen the clinical outcome of COVID-19. Despite this evidence supporting the high prevalence of co-infections among SARS-CoV-2 cases and its potentially substantial clinical impacts on COVID-19, existing data on co-infection remain limited by the low representation of the global population and the small number of viruses examined. To date, most studies have tested only a narrow spectrum of viruses using targeted qRT-PCR assays5,6, and only one reported on the co-infection rate among SARS-CoV-2 cases in the Southern Hemisphere7.

Here we measured the rate of respiratory viral co-infection among SARS-CoV-2 cases in Australia, determined by respiratory virome sequencing (excluding phages) using two commercial hybrid-capture sequencing platforms: (i) Virome Capture Sequencing (VirCapSeq), a collection of ~ 2 million oligonucleotide probes (70–120...
mers) targeting all known vertebrate-infecting viruses from Roche Sequencing; and (ii) Twist Respiratory Virus Panel, consisting of 41,047 probes (120 mers) targeting 29 human respiratory viruses representing six major pathogenic viral clades, from Twist Biosciences. Furthermore, we evaluated the feasibility of simultaneously performing SARS-CoV-2 whole genome sequencing (WGS) analysis using data generated from both methods. This demonstrated the utility of such a process, and greater WGS coverage achieved using hybrid-capture sequencing over existing amplicon-based procedures in situations where primer binding sites are abolished by genomic deletions.

Results

Rate of viral co-infection in Australian cases. We examined the respiratory virome of 92 SARS-CoV-2 cases who tested positive for SARS-CoV-2 RNA between March and May 2020 in New South Wales (NSW), Australia (Supplementary Table 1). The abundance of SARS-CoV-2 RNA in the respiratory specimens obtained from these cases was diverse, with qRT-PCR cycle threshold (Ct) values ranging between 13.3 and 39.7. This was equivalent to a viral load range between $1.4 \times 10^8$ copies/mL and less than 10 copies/mL (Supplementary Table 2). Sequencing all vertebrate-infectious viruses in these specimens using VirCapSeq hybrid-capture generated a total of 982 million raw reads, an average of 3.6 million adapter/host filtered reads per sample.

Overall, 47 species of viruses belonging to 17 different genera were detected with a minimum of 20 virus-classified reads per million (rpM; Supplementary File 1; Fig. 1a). None of these viruses were detected in the negative controls (SSC_1 and SSC_2), ruling out laboratory contamination or index switching as a source of spurious virus detection (Fig. S1a). SARS-CoV-2 reads above the positivity threshold were detected in 80% of samples (74/92), of which the highest Ct was 39.7. No SARS-CoV-2 reads were detected in the coronavirus-negative clinical controls (nCoV_neg_1 and _2; Fig. 1a). Among other respiratory viruses, picornaviruses (all of which were rhinoviruses) were detected in 5% (5/92) of cases and in nCoV_neg_1, and influenzavirus A in 2% (2/92) of cases. Consistent with our published data in other cohorts, the non-respiratory viruses detected included mammarenaviruses (41%), roseoloviruses (36%), alphapolyomaviruses (35%), papillomaviruses (20%), lymphocryptoviruses (12%), lentiviruses (9%), anelloviruses (3%), simplexviruses (3%), pestiviruses (3%), mastadenovirus (1%) and norovirus (1%). Overall, sequences of viruses other than SARS-CoV-2 were detected in 74% (68/92) of cases, and the rate of co-infection between SARS-CoV-2 and other respiratory viruses was 8% (7/92).
Validation using the Twist Respiratory Virus Panel. The VirCapSeq results were validated in 85 of the same pre-capture libraries, characterized in parallel using the Twist Respiratory Virus Panel that targets 29 human respiratory viruses. In total, 747 million raw reads were generated, averaging 7.4 million filtered reads per sample. Overall, 20 species of virus were detected, belonging to seven genera (Supplementary File 1; Fig. 1b). None of these viruses were detected in the negative controls (SSC_3 and SSC_4), ruling out laboratory contamination or index switching as a source of spurious virus detection (Fig. S1b). Some sequences in the negative controls were determined as false-positive hits to human adenovirus B, arising from alignment of reads to human genomic DNA sequences cloned within an adenovirus vector backbone (Fig. S1c). SARS-CoV-2 reads were detected in 95% (79/83) of cases, including 16/83 samples undetected using VirCapSeq. SARS-CoV-2 was absent in both coronavirus-negative controls. Consistent with VirCapSeq results, rhinovirus and influenzavirus were the only other respiratory viruses detected. Moreover, the Twist and VirCapSeq panels showed good concordance, with all samples positively identified for these viruses being detected on both platforms, with the exception of one additional rhinovirus-positive case detected by Twist (nCoV_235; Fig. 1a,b). We note that this discordant case was only marginally above the positivity threshold and reads covered < 5% of the reference genome (Supplementary Table 3). Other low-level positives included non-respiratory viruses: bocaparvoviruses, cytomegalovirus, flavivirus and roseolovirus (Fig. 1b).

Validation across multiple bioinformatics pipelines. The choice of de novo assembler can profoundly impact the analysis and interpretation of virome sequencing data12. To test whether the respiratory viral co-infections detected by VirCapSeq and Twist capture sequencing were reproducible using other pipelines that apply a different assembler or a k-mer approach, we compared IDseq results to outputs of VirMAP13 and OneCodex14 (Supplementary Table 4), respectively. Although the total number of taxonomically classified reads varied across the three pipelines for all respiratory viruses, IDseq and VirMAP were fully concordant for rhinovirus and influenzavirus positive samples. There were clear differences in the number samples identified as SARS-CoV-2 positive, OneCodex identifying the fewest despite reporting the highest total number of SARS-CoV-2 classified reads. IDseq was the most sensitive in identifying SARS-CoV-2 positive samples.

Complete genome coverage of co-infecting influenza virus. Unlike targeted qRT-PCR assays, virome capture sequencing can be used to determine the genome sequences of co-infecting viruses, informing investigations of virus transmission and evolution. We assessed the suitability of both panels for genome sequencing of co-infecting respiratory viruses. No clear difference was observed between VirCapSeq and Twist capture with respect to rhinovirus genome coverage. Breadth of coverage ranged between 24.3–97.4% and 14.7–75.2% of the reference genome, respectively in samples positive for rhinovirus on both platforms (Supplementary Table 3; Fig. S2). In contrast, VirCapSeq consistently achieved higher mean depth of coverage across most segments of the influenzavirus (Supplementary Table 5; Fig. S2). Nevertheless, Twist capture sequencing provided complete genome coverage of co-infecting influenzaviruses at high depth, sufficient to detect single nucleotide variants (SNVs) at the consensus level (Fig. 2). For both rhinoviruses and influenzaviruses, identified types were concordant between VirCapSeq and Twist panels. Therefore, both platforms were suitable for sequence determination of co-infecting respiratory viruses.

Figure 2. Full genome coverage of co-infecting influenzavirus. Coverage plot of sequence reads generated by Twist capture sequencing of the SARS-CoV-2 case specimen nCoV_240, aligned to the influenzavirus A reference genome across eight different segments (S1-S8). Depth represented as X fold coverage. Single nucleotide polymorphisms (SNPs) detected at positions across the genome are indicated in red, greater intensity of red indicates higher % frequency.
SARS-CoV-2 genome coverage. Viral WGS is being widely applied to study the transmission and evolution of SARS-CoV-2. Amplicon-based sequencing is currently used most frequently but has some limitations in scalability and reproducibility. Given the high sensitivity of the Twist panel for detecting SARS-CoV-2 reads even at viral loads near the qRT-PCR limit of detection (Fig. 3a), we investigated whether the Twist sequencing data provided sufficient coverage of SARS-CoV-2 for WGS analysis. The mean number of SARS-CoV-2 reads detected by the Twist was > tenfold higher than for VirCapSeq (Fig. S3). For samples quantitated with a Ct < 30 on qRT-PCR, Twist capture sequencing achieved a minimum > tenfold sequencing depth across > 90% the SARS-CoV-2 genome for 95% (57/60) of samples (Fig. 3b), and > 30-fold depth for 89% (53/60) of samples (Fig. S4). The highest Ct at which Twist provided > 90% coverage of > tenfold depth was 32.1. Even in the sample with the lowest viral load (Ct 39.7), 91% of genome was covered at 1X depth (Supplementary File 2).
Compared with the Twist platform, VirCapSeq did not achieve > 90% coverage even at 1X depth for all samples, except two with Ct < 20 (Fig. 3b; Fig. S4; Supplementary File 2). To test whether this low genome coverage was primarily due to the fewer number of SARS-CoV-2 reads detected compared to Twist, SARS-CoV-2 reads in the Twist dataset were sub-sampled to equal to that of VirCapSeq. Even after sub-sampling, Twist-enriched sequences maintained > 90% coverage of 10X depth across the SARS-CoV-2 genome in > 50% of the samples (Fig. S5). This demonstrated that the superior coverage of the SARS-CoV-2 genome achieved using Twist capture sequencing was only partially due to the higher number of SARS-CoV-2 reads. Coverage heterogeneity was the more important determinant, with reads being uniformly distributed across the SARS-CoV-2 genome in Twist samples but scattered unevenly for VirCapSeq (Fig. 3c). Therefore, while the VirCapSeq panel is suitable for detection of SARS-CoV-2 and coinfecting viruses, it is unsuitable for SARS-CoV-2 WGS.

Detection of inter-individual variation of SARS-CoV-2.

We evaluated the capacity to identify inter-individual genetic variation of SARS-CoV-2 from Twist sequencing data. Among 83 cases examined by Twist capture sequencing, the SARS-CoV-2 genomes of 48 cases were previously characterized from same samples through amplicon-based WGS on the Illumina platform. This confirmed the presence of inter-individual single nucleotide variants (SNVs) at the consensus level. Taking the amplicon-WGS data as the truth set, we assessed the sensitivity and precision of consensus sequence variants detected from Twist sequences. Overall, 338 consensus level SNVs were detected with 96% sensitivity and 98% precision, perfectly identified in 88% (42/48) of samples examined (Supplementary Table 6).

Detection of ORF8 deletion and validation using amplicon WGS. Multiple studies report major structural variations (SVs) in the SARS-CoV-2 genome, namely the 382 nt deletion in the open reading frame 8 (ORF8), associated with changes in the replicative fitness and milder infections. Therefore, we investigated if similar ORF8 deletions could be detected from the cases examined by Twist capture sequencing. We identified two cases with a common 328 bp deletion in ORF8 (nCoV_200 and nCoV_225). Providing further validation, the same deletion was detected in both cases through amplicon based WGS (Fig. 4; Fig. S6), using Oxford Nanopore Technology (ONT). Interestingly, in both cases, the deletion abolished a primer site, causing the failure of an adjacent amplicon (2.5 kb) and resulting in incomplete coverage of the SARS-CoV-2 genome when profiled by amplicon sequencing. By contrast, hybrid capture sequencing was able to achieve complete genome coverage. This demonstrates that Twist capture sequencing achieves sufficient coverage to reliably detect large deletions in the SARS-CoV-2 genome for clinical specimens of Ct < 30 and is more robust to large deletion or rearrangements in the genome, which can disrupt amplicon schemes.

Discussion

Determining the co-infection rate and consequent clinical impacts on COVID-19 is critical, particularly where therapeutic interventions for some coinfecting agents such as influenzavirus are available. In this study, we sequenced the respiratory virome using two hybrid-capture approaches and multiple taxonomic read classification pipelines, demonstrating an 8% rate of co-infection with other respiratory viruses among SARS-CoV-2 cases in Australia. This is less than half the rate reported in Northern California, but greater than the initial estimates...
from Wuhan (0.0%)\(^1\) and rates of viral co-infection reported from Chicago (1.6%)\(^20\), New York (2.0%)\(^21\), Singapore (1.4%)\(^22\), Barcelona (0.6%)\(^23\), and Turkey (2.0%)\(^24\). Furthermore, it is higher than the 4.6% of co-infection observed among 175 cases from the same region, diagnosed and tested during a similar time period using multiplex qRT-PCR\(^7\). Such high inter- and intra-regional variability warrants further investigation, particularly in developed countries with similar SARS-CoV-2 incidence to Australia. Recent data from Iran\(^4\) and Poland\(^24\) support higher mortality of COVID-19 patients as a result of respiratory viral co-infections.

Previous studies have reported co-infections between SARS-CoV-2 and common respiratory viruses including rhinovirus, influenza virus, metapneumovirus, parainfluenzavirus and respiratory syncytial virus\(^25\). In our results, co-infections between SARS-CoV-2 and rhinoviruses (6%) were predominant, lower for influenza viruses (2%). Interestingly, co-infection between SARS-CoV-2 and influenza was not observed in a recent report of Australian SARS-CoV-2 cases\(^6\). The case specimens examined in the present study were collected between March and May 2020, overlapping with the start of the Southern Hemisphere influenza season. In Australia, the flu season thus far has reported >90% reduction in incidence of influenza virus infections compared to the same period in 2019 as a result of social distancing and mandatory quarantine measures applied during the COVID-19 pandemic\(^26\). Therefore, the significant reduction in circulating influenza virus may have contributed to its low co-infection rate with SARS-CoV-2.

Our study has several limitations. All specimens examined in this study were freeze-thawed twice before library preparation. This may have prevented detection of viruses that were originally at very low titer. Therefore, the actual rate of co-infection may exceed 8%. To eliminate this potential in future analyses, double-stranded cDNA should be generated on the same day as SARS-CoV-2 qRT-PCR, from the same nucleic acid extracts avoiding freeze-thaw. A key limitation of our analysis was the lack of clinical metadata, precluding examination of potential associations between respiratory viral co-infection with SARS-CoV-2 and clinical outcomes of COVID-19. Although comparable to other co-infection studies to date, our sample size was small and included only a single timepoint for each case. Nevertheless, this represents the largest metagenomic sequencing study to date, examining co-infections between SARS-CoV-2 and other respiratory viruses.

There is growing appreciation for SARS-CoV-2 WGS as an essential tool to investigate the transmission and evolution of SARS-CoV-2, critical for research and public health responses to COVID-19\(^15–17,27–31\). Existing WGS approaches can be divided into two main categories: 1. Amplicon sequencing; and 2. Hybrid-capture sequencing using SARS-CoV-2-specific probes. Neither are capable of simultaneously detecting co-infecting viruses. Our analysis of the SARS-CoV-2 genome using Twist-enriched sequenced demonstrated high breadth and depth of coverage for samples with Ct < 30, sufficient for downstream analysis of SNV, indels and SVs. This was despite using single-end sequence data. Hence, even greater confidence in variant calling can be achieved using paired-end sequencing. Overall, target enrichment sequencing using the Twist Respiratory Virus Panel offers dual-functionality, providing effective characterization of co-infecting respiratory viruses and the full genome of the SARS-CoV-2, simultaneously.

Unlike amplicon sequencing, Twist hybrid-capture does not require generation of SARS-CoV-2 amplicons. This significantly reduces processing time and manual handling, lowering the risk of cross contamination. Using the Twist's fast hybridization and multiplexed capture workflow, libraries ready for high throughput sequencing can be constructed from clinical specimen extracts in < 8 h. Although the amplicon approach can also construct libraries within a similar timeframe, in our experience of using two published amplicon WGS methods\(^15,28,32\), generating amplicons often took longer than anticipated due to certain parts of the genome amplifying poorly, requiring continuous optimization. Furthermore, in this study, all libraries were hybridized with Twist probes for 2 h. However, this can be reduced to 30 min with minimal loss in capture efficiency.

The default protocol for Twist hybrid-capture supports multiplexing up to 8 libraries (8-plex) per capture hybridization, combining libraries by equal mass to make up 1.5 μg of total DNA, or up to 4 μg total without compromising the efficacy of target enrichment. In this study, we performed Twist capture on libraries pooled up to 20-plex, whilst still maintaining the 4 μg total DNA limit. This highly multiplexed sample processing significantly reduced processing time, labor and cost per sample. Current per-sample cost of Twist Respiratory Virus Panel in a 20-plex sample format is $25 USD. This compares favorably with the cost of VirCapSeq (~ $23 USD per sample), particularly given its advantages in sensitivity and genome coverage of SARS-CoV-2.

Taken together, we provide a practical and cost-effective strategy for characterizing both respiratory viral co-infections and the full SARS-CoV-2 genome simultaneously, from clinical specimens with Ct < 30 or viral load > 3,000 copies/mL. We also recommend IDseq as the preferred pipeline for taxonomic classification of viral sequences in SARS-CoV-2 specimens, based on its high sensitivity for SARS-CoV-2 and other respiratory viruses, ease of use, and minimal requirements in terms of infrastructure and bioinformatic expertise. We envision broad application of our approach across research and clinical settings.

**Methods**

**Clinical samples and SARS-CoV-2 qRT-PCR.** Respiratory specimens of SARS-CoV-2 cases (adults) in NSW diagnosed between March and May 2020 were obtained from at the Prince of Wales Hospital in Randwick, Sydney, Australia. Ethical approval and informed consent waiver was received from the South Eastern Sydney Local Health District Human Research Ethics Committee (2020/ETH02639). All methods were performed in accordance with the relevant guidelines and regulations. Prior to this study, samples were freeze-thawed twice and stored at ~ 80 °C following diagnostic testing at the NSW Health Pathology East Serology and Virology Division (SaViD). In total, 92 nasopharyngeal swabs suspended in Viral Transport Media (VTM) were selected for virome capturing sequencing, all positive for a combination of four SARS-CoV-2 target genes (RdRp, S, N and E) on the Allplex SARS-CoV-2 qRT-PCR Assay (Seegeen, Seoul, Korea). The approximate copy number of SARS-CoV-2 RNA was calculated by plotting the Ct against a standard curve built from tenfold serial dilution of a
quantified N-gene plasmid control, developed inhouse. To use as controls, two nasopharyngeal swabs confirmed negative for SARS-CoV-2 from the same diagnostic laboratory and two negative controls prepared from purified Salmon Sperm DNA (15632-011; Thermo Fisher Scientific, MA, USA) were also sequenced.

**Total nucleic acid extraction, cDNA synthesis and library prep.** Total nucleic acid was extracted from 200 µL of swab suspension in VTM, using the AllPrep PowerVirial DNA/RNA kits (Qiagen, Hilden, Germany) with bead-beating and phenol/chloroform, following manufacturer’s protocol. Using Supernscript III (Thermo Fisher Scientific) and Klenow Fragment (NEB, MA, USA) with random hexamers, the RNA portion was converted into double-stranded cDNA. Illumina sequencing libraries were prepared from 1ug of double-stranded DNA/cDNA mixture, using the KAPA Hyper Plus (Roche, Basel, Switzerland) kit with Unique Dual-Index adapters. Libraries were quantified by picogreen (Thermo Fisher Scientific) and the size distribution of library fragments were measured on the LabChip GX Touch 24 (Perkin Elmer, MA, USA) bioanalyzer.

**Target enrichment sequencing.** For VirCapSeq hybrid-capture, individual libraries (92 cases, 2 coronavirus-negative controls and 2 salmon sperm DNA controls) were combined by equal mass into two capture pools (48-plex) and hybridized to probes (VirCapSeq-VERT design; Roche) for 16 h as previously described14, following the SeqCap EZ HyperCap Workflow v2.3 (Roche). For hybrid-capture using the Twist Respiratory Virus Research Panel (103067; Twist Biosciences, San Francisco, CA), 87 libraries (83 cases, 2 coronavirus-negative controls and 2 salmon sperm DNA controls) were combined by equal mass into five capture pools (16- to 20-plex; average 17-plex). Pools were hybridized to probes for 2 h, following the Fast Hybridization Workflow (Twist Biosciences). VirCapSeq and Twist post-capture library pools were PCR amplified 16 cycles and single-ended sequenced (1 × 100 bp) separately, up to 96 barcoded libraries maximum per lane of a NovaSeq 6000 S1 flowcell (Illumina, San Diego, CA) at the Ramaciotti Centre for Genomics (UNSW Sydney, Australia).

**Taxonomic classification of metagenomic reads.** By default, taxonomic classification of viral reads in all samples was achieved using IDseq (v4.0)35, a cloud-based, open-source bioinformatics pipeline for metagenomic sequencing data. Raw fastq files were uploaded to the IDseq portal (https://idseq.net) using the Amazon Web Services (AWS) Command Line Interface. All IDseq scripts and user instructions are available at https://github.com/chanzuckerberg/idseq-dag. In brief, adapter and human host sequences were filtered, and remaining short-read sequences were aligned to the NCBI nucleotide (nt) and nonredundant protein (nr) databases ([ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/) using GSNAP25 and RAsearch25, respectively. Putative accessions were assigned to each read using the NCBI accession2taxid database ([ftp://ftp.ncbi.nih.gov/pub/taxonomy/ accession2taxid] and a BLAST + (v 2.6.0)46 database. In parallel, short reads were de novo assembled into contigs using SPAdes37. Raw reads were mapped back to the resulting contigs using Bowtie228, to identify the contig to which they belong. Finally, each contig was aligned to the set of possible accessions represented by the BLAST database, to improve the specificity of alignments to all the underlying reads. Only viruses detected at ≥ 20 rpM based on nt alignments (NT rPM) were deemed positive and included in heatmaps generated using heatmapper39.

For the comparative analysis between IDseq and other bioinformatic pipelines, taxonomic read classification summaries were generated using VirMAP13 and OneCodex14. VirMAP was installed and run on the National Computing Infrastructure (NCI) HPC Gadi with modifications described at https://github.com/rsotone/virma. OneCodex is a premium cloud-based pipeline, for which raw fastq files generated using Twist hybrid-capture sequencing were uploaded and summary reports downloaded using the web browser interface (https://app.onecodex.com/).

**Virus genome assembly, coverage analysis and variant detection.** For samples containing sequences corresponding to rhinoviruses/enteroviruses and influenzaviruses, host-filtered sequences from IDseq were mapped to their respective reference genome sequence obtained from the NCBI database using minimap2 (v2.17-r941)40. Coverage statistics and SNV reports were generated from the sorted bam file using qualimap (v2.2.2-dev)41 and freebayes (v1.3.2-dirty)42, respectively. Genome assemblies and coverage statistics were also generated in Geneious Prime (v2020.2.2; Biomatters Ltd.)43 for supplementary tables and figures.

For all SARS-CoV-2 genome assemblies, host-filtered reads were aligned to the Wuhan-Hu-1 reference genome (MN908947.3) using bwa mem (0.7.12-r1039)34, with only MapQ = 60 alignments retained. Per-base coverage was calculated at each position in the SARS-CoV-2 genome using bedtools coverage (v2.25.0)35. Coverage breadth (fraction of positions covered) was calculated at a range of different minimum depths (≥ 1X, 2X, 5X, 10X, 20X, 30X, 50X, 100X). For samples where SARS-CoV-2 was detected at Ct < 30, SNVs were detected using samttools mpileup (v1.9.4)36 and Varscan2 mpileup2snv (v2.4.3)37. SNVs with a minimum read-count frequency ≥ 80% were retained as consensus SNVs. Variant detection performance was evaluated by comparing detected consensus SNVs in Twist capture samples to matched specimens analyzed elsewhere38 by amplicon-based WGS (n = 48). Validation of ORF8 deletions by amplicon-based WGS using the ONT platform was performed as previously described15 and alignment of reads across the deletion site was visualized in Integrative Genomics Viewer (IGV; v2.8.4)38.

**Data availability**

All de-identified metagenomic sequencing data (raw and processed fastq files) will be made publicly available in time for publication.
References


**Acknowledgements**


**Author contributions**


**Competing interests**

The authors declare no competing interests.

**Additional information**

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-83642-x.

Correspondence and requests for materials should be addressed to W.D.R.

Reprints and permissions information is available at www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021