JHU THESIS TEMPLATE TITLE

by

Yunfan Fn

A dissertation submitted to Johns Hopkins University
in conformity with the requirements for the degree of
Doctor of Philosophy

Baltimore, Maryland
August, 2022

© 2022 by Yunfan Fan
All rights reserved

Abstract

While next generation sequencing (NGS) has enabled massively parallel DNA sequencing for lower and lower cost, the development of third generation nanopore sequencing offers several key advantages over older sequencing methods. Nanopore sequencers are pocket-sized, making them orders of magnitude cheaper than the next most affordable alternative and the ideal option for wide deployment. They are capable of providing data in real-time, saving valuable hours before data analysis can begin. Additionally, they are able to sequence reads several thousand basepairs long, as opposed to the hundreds of basepairs NGS platforms are capable of, and they embed base modification data without the need for specific treatment beforehand. Given these advantages, in this thesis I examine the application of nanopore sequencing to the study of human pathogens.

First, we use nanopore sequencing to characterize anti-microbial resistance (AMR) in forty clinical isolates. We analyzed real-time data to quickly identify AMR genes, assembled genomes to identify chromosomal mutations, and used short-read sequencing data to correct the errors in the assemblies. With sequencing data, we found that time to effective antibiotic therapy could be shortened by as much as 20 hours compared to standard antimicrobial

susceptibility testing (AST).

Second, we leverage the long reads of nanopore sequencing to assemble the genome of a pathogenic yeast, /textitCandida nivariensis. Previous efforts to assemble this yeast genome relied solely on NGS data, resulting in a highly fragmented genome. Using nanopore data, we achieve a much higher contiguity, capture previously missing portions of the genome. Furthermore, we demonstrate that our more contiguous genome can be used to better study long and repetative genes, such as those involved in pathogenticity to humans.

Third, we use the base modification information embedded in nanopore sequencing data to call methylation in metagenomic assemblies. These calls enable the binning of metagenomic contigs according to methylation signature without the need to collect additional data. We demonstrate the efficacy of this method on a synthetic community sample, a simple two-bacteria system, and a clinical sample with matched proximity ligation binning data.

These applications of nanopore sequencing demonstrate its potential and its utility for all fronts of pathogen genomics research.

Thesis Committee

Primary Readers

Dr. Winston Timp (Primary Advisor)
Associate Professor
Department of Biomedical Engineering
Department of Molecular Biology and Genetics
Johns Hopkins University School of Medicine

Dr. Patricia Simner
Associate Professor
Department of Pathology
Johns Hopkins University School of Medicine

Dr. Steven Salzberg
Bloomberg Distinguished Professor
Department of Computer Science
Johns Hopkins University Whiting School of Engineering
Department of Biomedical Engineering
Johns Hopkins University School of Medicine
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Alternate Readers

First Lastname
Professor
Department of ChangeMe
Johns Hopkins Bloomberg School of Public Health

First Lastname
Assistant Professor
Department of ChangeMe
Johns Hopkins Bloomberg School of Public Health

Acknowledgments

I have tremendous gratitude
to those people,
numerous and uncountable,
who have contributed,
directly or in subtler ways,
to this work.

Some of them are listed here.

To my advisor, Winston: I remember writing to you as a sophomore in college many years ago, asking to do research in your brand new lab, which at the time was but a few months old. Back then, I hardly knew what research was and had no relevant skills or credentials to offer, only my time and my interest to learn. Over these years I've learned so much from you, and will always be grateful to you for building the place where I was able to grow.

To my thesis committee, Trish and Steven: Thank you for your patient guidance, encouragement, advice, and for helping me to keep an eye on the bigger picture.

To the @yfan arc of the #core channel - @isac, @brochael, @shao, @gilfunk,

@narley, @broham, @gmoney, @Brittany, @sherbear, @Sam Sholes, @Paul Hook, @amymeltzer39, and @alice: Thank you for those times when you patiently watched over me as I learned new lab techniques, answered my dumb questions, and generally saved me from my own buffoonery. Thank you even more for commiserating with me as we struggled together through the singular challenges of research, and celebrating the equally singular triumphs.

To the crew that moved me into Boonique, and Charles, and Charlotte, and Sven, and Manolo: Thanks for being there.

To mom and dad, and family further away: It was your labor that first cultivated my growth. Accomplishments in my name are as much yours as they are mine. I flourish for you.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Introduce your thesis (Aardvark, 1900)

References

Aardvark, A. A. (1900). “Article title”. In: *Journal One* 1.1, pp. 1–8.

Chapter 2

Genome assembly of Candida nivariensis

Portions of this chapter originally appeared in:

Fan Y, Gale AN, Bailey A, Barnes K, Colotti K, Mass M, et al. Genome and transcriptome of a pathogenic yeast, Candida nivariensis. G3 Genes | Genomes | Genetics. 2021;11. doi:10.1093/g3journal/jkab137

2.1 Abstract

We present a highly contiguous genome and transcriptome of the pathogenic yeast, Candida nivariensis. We sequenced both the DNA and RNA of this species using both the Oxford Nanopore Technologies and Illumina platforms. We assembled the genome into an 11.8Mb draft composed of 16 contigs with an N50 of 886 Kb, including a circular mitochondrial sequence of 28 Kb. Using direct RNA nanopore sequencing and Illumina cDNA sequencing, we constructed an annotation of our new assembly, supplemented by lifting over genes from Saccharomyces cerevisiae and Candida glabrata.

2.2 Introduction

For immunocompromised hosts, opportunistic infections caused by drug-resistant fungi of the Candida genus are a major source of morbidity and mortality (Borman et al., 2008). In particular, Candida nivariensis, a close relative to Candida glabrata, has emerged in recent years as especially resistant to antifungal therapies (Borman et al., 2008). However, due to its phenotypic similarities to C. glabrata, C. nivariensis is generally underidentified and easily misdiagnosed, and currently, only molecular approaches can distinguish the two (Aznar-Marin et al., 2016), spurring whole-genome sequencing studies on the clade (Gabaldón et al., 2013).

Accurate assembly of repetitive genomic regions is crucial for understanding genetic diversity and virulence in pathogenic species. Fungal pathogens have long been known to exhibit a high degree of genome plasticity to enhance fitness in various environments (Croll, Zala, and McDonald, 2013; Ford et al., 2015; López-Fuentes et al., 2018; Carreté et al., 2019; Todd et al., 2019). Repetitive subtelomeric regions in particular play a crucial role in virulence for many pathogenic organisms (Barry et al., 2003; De Las Peñas et al., 2003). Many yeasts' subtelomeric regions contain and regulate the expression of genes crucial for biofilm formation, carbohydrate utilization, and cellular adhesion (Naumov, Naumova, and Louis, 1995; De Las Peñas et al., 2003; Iraqui et al., 2005). These gene families often undergo rapid evolution through changes in copy number and sequence through either SNPs or indels (Carreto et al., 2008; Brown, Murray, and Verstrepen, 2010; Anderson et al., 2015). However, these subtelomeric regions remain one of the most difficult sections of the genome to

accurately assemble due to their repetitive nature and high sequence similarity between genes, making genetic analysis cumbersome (Brown, Murray, and Verstrepen, 2010).

One of the gene families of great interest to the pathogenic yeast field are the GPI-anchored cell wall proteins. This protein family includes many genes that encode for adhesion proteins that are found in various members of the Candida genus, and play a key role in pathogenicity, being involved in regulation of biofilm formation, cell-to-cell contact, and host-pathogen interactions (Timmermans et al., 2018; McCall et al., 2019). With the many roles these genes play in infection, the accurate identification and understanding of the genetic variation of these genes vital to combating fungal pathogens.

Unfortunately, like many eukaryotic pathogens, the current reference genome for C. nivariensis (GenBank: GCA_001046915.1) is highly fragmented. Constructed from sequencing of strain CBS9983, the reference genome consists of 123 contigs with an N50 of 248Kb (Gabaldón et al., 2013), meaning that at least half of the total genome length is contained in contigs 248Kb or longer. This is typical of genomes assembled from limited short-read sequencing data; though short reads are highly accurate, assembling them into contiguous genomes is challenging depending on the size and complexity of the genome. Such short read assemblies have limited utility since large scale variants, repetitive regions, and genome structure remain difficult to elucidate, though they are often involved in the genome plasticity of pathogenic yeasts (Carreté et al., 2018). In contrast, long-read sequencing data has been shown to produce much more contiguous assemblies, and have been crucial

in sequencing through large repetitive regions, as well as assessing structural variants. However, read accuracy on the ONT platform in particular ranges from 86

Having a genome alone is not enough; we need to annotate it with genes and other functional elements for the genome to be of greatest use. Knowledge of gene loci is critical to constructing phylogenetic relationships between organisms, and to studying the functional implications of variants, both common uses of reference genomes. While model-based, purely computational gene predictors can be highly accurate in bacteria, gene sparsity and intronic regions make this task more difficult in eukaryotes (Salzberg, 2019). For improved annotations, some RNA-seq information is required (Salzberg, 2019).

Here, as part of our newly developed Methods in Nucleic Acid Sequencing university course, we used a hybrid approach, applying long-read nanopore sequencing to assemble a highly contiguous genome of C. nivariensis, followed by short-read sequencing to polish or correct errors in our assembly. We followed this by a combination of nanopore direct RNA sequencing as well as short-read RNA-seq to annotate our assembly. By combining this data with liftover of annotations from evolutionary "cousins" of nivariensis, we have generated a new and annotated reference genome for the community.

```
what <- "figure"
figname <- "asms"
title <- "Characteristics of the JHU_Cniv_v1"
desc <- "{\\bf (A)} Cumulative lengths of the 50 longest sequences in our assembly an
shortname <- "fig:asms"
```

```
where <- "[!ht]"


figstr <- paste0('\\begin{', what ,'}', where ,'
\\centering
\\includegraphics[width = 1\\linewidth,keepaspectratio]{figure/', figname, '.pdf}
\\caption[', title ,']{{\\bf ', title  ,'.} ', desc ,' }
\\label{', shortname ,'}
\\end{', what ,'}
')
cat(figstr)
```

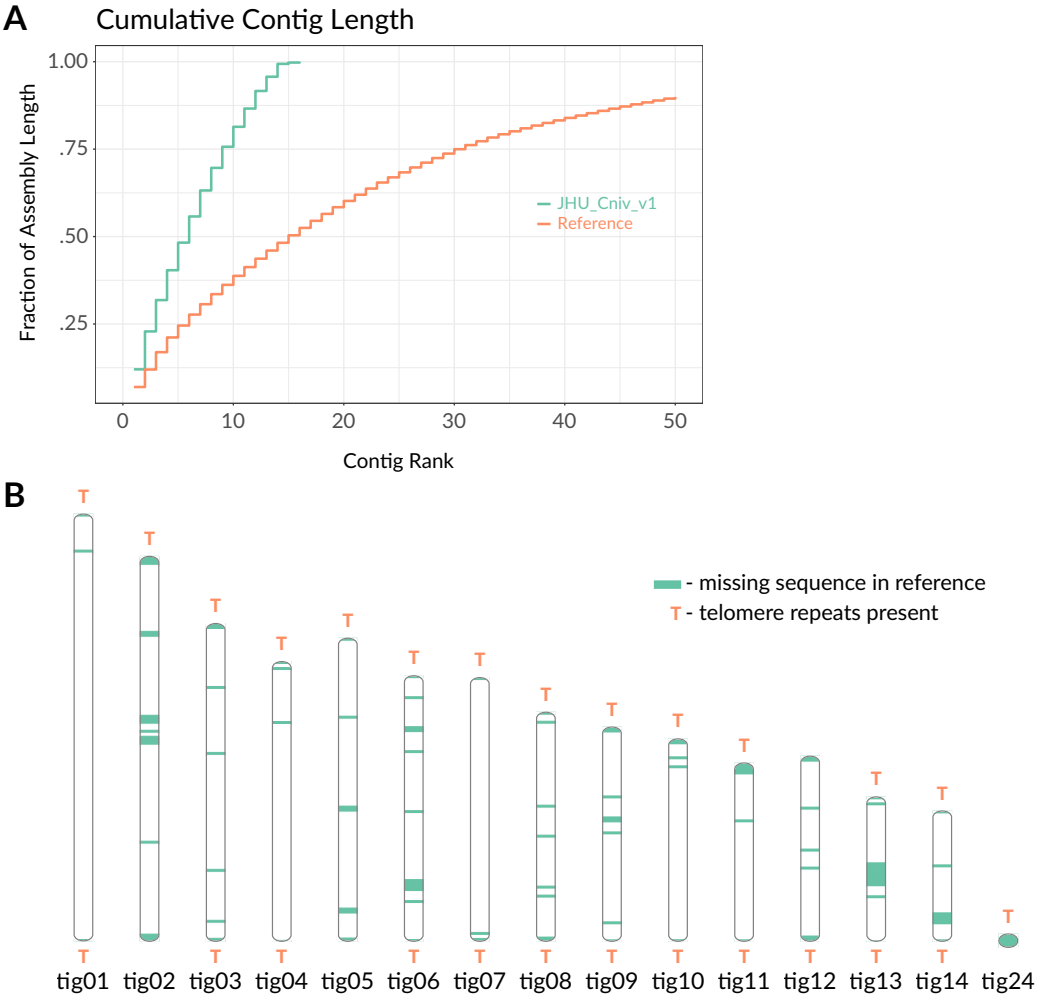

Figure 2.1: Characteristics of the JHU$_C niv_v 1.$(A)$Cumulativelengthsofthe50longestsequencesinourassem

References

Borman, Andrew M, Rebecca Petch, Christopher J Linton, Michael D Palmer, Paul D Bridge, and Elizabeth M Johnson (2008). "Candida nivariensis, an emerging pathogenic fungus with multidrug resistance to antifungal agents". en. In: *J. Clin. Microbiol.* 46.3, pp. 933–938.

Aznar-Marin, Pilar, Fátima Galan-Sanchez, Pilar Marin-Casanova, Pedro García-Martos, and Manuel Rodríguez-Iglesias (2016). "Candida nivariensis as a New Emergent Agent of Vulvovaginal Candidiasis: Description of Cases and Review of Published Studies". en. In: *Mycopathologia* 181.5-6, pp. 445–449.

Gabaldón, Toni, Tiphaine Martin, Marina Marcet-Houben, Pascal Durrens, Monique Bolotin-Fukuhara, Olivier Lespinet, Sylvie Arnaise, Stéphanie Boisnard, Gabriela Aguileta, Ralitsa Atanasova, Christiane Bouchier, Arnaud Couloux, Sophie Creno, Jose Almeida Cruz, Hugo Devillers, Adela Enache-Angoulvant, Juliette Guitard, Laure Jaouen, Laurence Ma, Christian Marck, Cécile Neuvéglise, Eric Pelletier, Amélie Pinard, Julie Poulain, Julien Recoquillay, Eric Westhof, Patrick Wincker, Bernard Dujon, Christophe Hennequin, and Cécile Fairhead (2013). "Comparative genomics of emerging pathogens in the Candida glabrata clade". en. In: *BMC Genomics* 14, p. 623.

Croll, Daniel, Marcello Zala, and Bruce A McDonald (2013). "Breakage-fusion-bridge cycles and large insertions contribute to the rapid evolution of accessory chromosomes in a fungal pathogen". en. In: *PLoS Genet.* 9.6, e1003567.

Ford, Christopher B, Jason M Funt, Darren Abbey, Luca Issi, Candace Guiducci, Diego A Martinez, Toni Delorey, Bi Yu Li, Theodore C White, Christina Cuomo, Reeta P Rao, Judith Berman, Dawn A Thompson, and Aviv Regev (2015). "The evolution of drug resistance in clinical isolates of Candida albicans". en. In: *Elife* 4, e00662.

López-Fuentes, Eunice, Guadalupe Gutiérrez-Escobedo, Bea Timmermans, Patrick Van Dijck, Alejandro De Las Peñas, and Irene Castaño (2018). "Candida glabrata's Genome Plasticity Confers a Unique Pattern of Expressed Cell Wall Proteins". en. In: *J Fungi (Basel)* 4.2.

Carreté, Laia, Ewa Ksiezopolska, Emilia Gómez-Molero, Adela Angoulvant, Oliver Bader, Cécile Fairhead, and Toni Gabaldón (2019). "Genome Comparisons of Candida glabrata Serial Clinical Isolates Reveal Patterns of Genetic Variation in Infecting Clonal Populations". en. In: *Front. Microbiol.* 10, p. 112.

Todd, Robert T, Tyler D Wikoff, Anja Forche, and Anna Selmecki (2019). "Genome plasticity in Candida albicans is driven by long repeat sequences". en. In: *Elife* 8.

Barry, J D, M L Ginger, P Burton, and R McCulloch (2003). "Why are parasite contingency genes often associated with telomeres?" en. In: *Int. J. Parasitol.* 33.1, pp. 29–45.

De Las Peñas, Alejandro, Shih-Jung Pan, Irene Castaño, Jonathan Alder, Robert Cregg, and Brendan P Cormack (2003). "Virulence-related surface glycoproteins in the yeast pathogen Candida glabrata are encoded in subtelomeric clusters and subject to RAP1- and SIR-dependent transcriptional silencing". en. In: *Genes Dev.* 17.18, pp. 2245–2258.

Naumov, G I, E S Naumova, and E J Louis (1995). "Genetic mapping of the alpha-galactosidase MEL gene family on right and left telomeres of Saccharomyces cerevisiae". en. In: *Yeast* 11.5, pp. 481–483.

Iraqui, Ismail, Susana Garcia-Sanchez, Sylvie Aubert, Françoise Dromer, Jean-Marc Ghigo, Christophe d'Enfert, and Guilhem Janbon (2005). "The Yak1p kinase controls expression of adhesins and biofilm formation in Candida glabrata in a Sir4p-dependent pathway". en. In: *Mol. Microbiol.* 55.4, pp. 1259–1271.

Carreto, Laura, Maria F Eiriz, Ana C Gomes, Patrícia M Pereira, Dorit Schuller, and Manuel A S Santos (2008). "Comparative genomics of wild type yeast strains unveils important genome diversity". en. In: *BMC Genomics* 9, p. 524.

Brown, Chris A, Andrew W Murray, and Kevin J Verstrepen (2010). "Rapid expansion and functional divergence of subtelomeric gene families in yeasts". en. In: *Curr. Biol.* 20.10, pp. 895–903.

Anderson, Matthew Z, Lauren J Wigen, Laura S Burrack, and Judith Berman (2015). "Real-Time Evolution of a Subtelomeric Gene Family in Candida albicans". en. In: *Genetics* 200.3, pp. 907–919.

Timmermans, Bea, Alejandro De Las Peñas, Irene Castaño, and Patrick Van Dijck (2018). "Adhesins in Candida glabrata". en. In: *J Fungi (Basel)* 4.2.

McCall, Andrew D, Ruvini U Pathirana, Aditi Prabhakar, Paul J Cullen, and Mira Edgerton (2019). "Candida albicans biofilm development is governed by cooperative attachment and adhesion maintenance proteins". en. In: *NPJ Biofilms Microbiomes* 5.1, p. 21.

Carreté, Laia, Ewa Ksiezopolska, Cinta Pegueroles, Emilia Gómez-Molero, Ester Saus, Susana Iraola-Guzmán, Damian Loska, Oliver Bader, Cecile Fairhead, and Toni Gabaldón (2018). "Patterns of Genomic Variation in the Opportunistic Pathogen Candida glabrata Suggest the Existence of Mating and a Secondary Association with Humans". en. In: *Curr. Biol.* 28.1, 15–27.e7.

Salzberg, Steven L (2019). "Next-generation genome annotation: we still struggle to get it right". en. In: *Genome Biol.* 20.1, p. 92.

Chapter 3

Discussion and Conclusion

Discuss and conclude your thesis (Abramson, Barbie, and Rider, 1900)

References

Abramson, A. A., B. B. Barbie, and C. C. Rider (1900). "Article title". In: *Journal Three* 1.1, pp. 192–244.

John Doe

Resumé title

Some quote

Education

year–year **Degree**, *Institution*, City, *Grade*.
Description

year–year **Degree**, *Institution*, City, *Grade*.
Description

Master thesis

title *Title*

supervisors Supervisors

description Short thesis abstract

Experience

Vocational

year–year **Job title**, *Employer*, City.
General description no longer than 1–2 lines.
Detailed achievements:

- Achievement 1;
- Achievement 2, with sub-achievements:
  - Sub-achievement (a);
  - Sub-achievement (b), with sub-sub-achievements (don't do this!);
    - Sub-sub-achievement i;
    - Sub-sub-achievement ii;
    - Sub-sub-achievement iii;
  - Sub-achievement (c);
- Achievement 3.

year–year **Job title**, *Employer*, City.
Description line 1
Description line 2

Miscellaneous

street and number – postcode city – country
+1 (234) 567 890 • *+2 (345) 678 901* • *+3 (456) 789 012*
john@doe.org • *www.johndoe.com* • *john.doe* • *jdoe*
jdoe • *additional information*

year–year **Job title**, *Employer*, City.
Description

Languages

Language 1 Skill level *Comment*
Language 2 Skill level *Comment*
Language 3 Skill level *Comment*

Computer skills

category 1 XXX, YYY, ZZZ
category 2 XXX, YYY, ZZZ
category 3 XXX, YYY, ZZZ
category 4 XXX, YYY, ZZZ
category 5 XXX, YYY, ZZZ
category 6 XXX, YYY, ZZZ

Interests

hobby 1 Description
hobby 2 Description
hobby 3 Description

Extra 1

- Item 1
- Item 2
- Item 3. This item is particularly long and therefore normally spans over several lines. Did you notice the indentation when the line wraps?

Extra 2

- Item 1
- Item 2
- Item 3
- Item 4
- Item 5[3]
- Item 6. Like item 3 in the single column list before, this item is particularly long to wrap over several lines.

References

Category 1

- Person 1
- Person 2
- Person 3

Category 2

Amongst others:

- Person 1, and
- Person 2

(more upon request)

All the rest & some more

That person, and **those** also (all available upon request).

Publications

[1] John Doe. Title, year.

street and number – postcode city – country
+1 (234) 567 890 • *+2 (345) 678 901* • *+3 (456) 789 012*
john@doe.org • *www.johndoe.com* • *john.doe* • *jdoe*
jdoe • *additional information*

[2] John Doe. Title, year.

[3] John Doe and Author 1. *Title*. Publisher, edition edition, year.

[4] John Doe and Author 2. *Title*. Publisher, edition edition, year.

[5] John Doe and Author 3. Title, year.

Company Recruitment team January 01, 1984
Company, Inc.
123 somestreet
some city

Dear Sir or Madam,

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis ullamcorper neque sit amet lectus facilisis sed luctus nisl iaculis. Vivamus at neque arcu, sed tempor quam. Curabitur pharetra tincidunt tincidunt. Morbi volutpat feugiat mauris, quis tempor neque vehicula volutpat. Duis tristique justo vel massa fermentum accumsan. Mauris ante elit, feugiat vestibulum tempor eget, eleifend ac ipsum. Donec scelerisque lobortis ipsum eu vestibulum. Pellentesque vel massa at felis accumsan rhoncus.

Suspendisse commodo, massa eu congue tincidunt, elit mauris pellentesque orci, cursus tempor odio nisl euismod augue. Aliquam adipiscing nibh ut odio sodales et pulvinar tortor laoreet. Mauris a accumsan ligula. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Suspendisse vulputate sem vehicula ipsum varius nec tempus dui dapibus. Phasellus et est urna, ut auctor erat. Sed tincidunt odio id odio aliquam mattis. Donec sapien nulla, feugiat eget adipiscing sit amet, lacinia ut dolor. Phasellus tincidunt, leo a fringilla consectetur, felis diam aliquam urna, vitae aliquet lectus orci nec velit. Vivamus dapibus varius blandit.

Duis sit amet magna ante, at sodales diam. Aenean consectetur porta risus et sagittis. Ut interdum, enim varius pellentesque tincidunt, magna libero sodales tortor, ut fermentum nunc metus a ante. Vivamus odio leo, tincidunt eu luctus ut, sollicitudin sit amet metus. Nunc sed orci lectus. Ut sodales magna sed velit volutpat sit amet pulvinar diam venenatis.

Albert Einstein discovered that $e = mc^2$ in 1905.

$$e = \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n$$

Yours faithfully,

John Doe

Attached: curriculum vitæ

John Doe
street and number – postcode city – country
+1 (234) 567 890 • *+2 (345) 678 901* • *+3 (456) 789 012*
john@doe.org • *www.johndoe.com* • *john.doe* • *jdoe*
jdoe • *additional information*