



"Looking for that perfect fit: Overcoming roadblocks to an optimal sequencing strategy with biocontrol agent *Nesidiocoris tenuis* as a case study" by Kim Ferguson is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/)

Looking for that perfect fit:

Overcoming roadblocks to an optimal sequencing strategy with biocontrol agent *Nesidiocoris tenuis* as a case study

Kim Ferguson

PhD Candidate
Wageningen University

[@kfergy](https://twitter.com/kfergy)



Given at the Royal Entomological Society's 2nd
Meeting of the Insect Genomics Special Interest
Group, 16th May 2017



This presentation was given at the Royal Entomological Society's 2nd Meeting of the Insect Genomics Special Interest Group, 16th May 2017. After slight adaptation, it was published online on 24th May, 2017. Please note photo credits as well as acknowledgements that occur on slides 17 and 28, and enjoy! This presentation is licenced as CC by 4.0, Attribution 4.0 International. Refer to <https://creativecommons.org/licenses/by/4.0/> for more info, but most important is to attribute the original to me.

BINGO: Breeding Invertebrates for Next Generation BioControl



- Key idea: Improvement of biocontrol agents via natural genetic variation
- Network of universities, research institutes, and companies
 - 13 PhDs
 - 24 senior researchers
 - 12 partners from Netherlands, Germany, France, Spain, Czech Republic, Austria, Switzerland, Greece and Portugal



2



The network that I operate in is called BINGO, short for Breeding Invertebrates for Next Generation BioControl. Biocontrol, also known as biological control as well as the use of beneficial insects in specific cases, is the use of one organism to control a pest population. While there are several uses for biocontrol, including human health and conservation, the usage that BINGO hopes to improve is its use in food crops. Biocontrol offers a safe and effective alternative to pesticides, when used correctly. The key idea of the BINGO project is the improvement of existing biocontrol agents via natural genetic variation. Additionally, we use next generation sequencing technology to inform our understanding (such as through genome assembly). We are a network of universities, research institutes, and companies, and seek to train 13 PhDs. I'm one of them!

My project:

Assemble the genome of three important biocontrol agents and compare the genetic variation between commercial strains and native populations

- genetic bottlenecks
- lab adaptation
- coding and non-coding regions
- aimed at making genome research attractive to biocontrol

Trichogramma brassicae



Sophie Chattington, Uni Bremen

Amblyseius swirskii



Angeliki Paspali, IVIA

Nesidiocoris tenuis




Milena Chinchilla-Ramirez, IVIA



3



My project in particular is to assemble the genome of three important biocontrol agents and compare the genetic variation between commercial strains and native populations. I am especially concerned about potential genetic bottlenecks that occur at the moment of capture, followed by years of lab adaptation. I will be looking into coding and non-coding regions, and this project is aimed at making genome research attractive to biocontrol practitioners. The three species I work on are in other projects within BINGO, so I also collaborate with three other PhDs. The work on *Trichogramma brassicae* is with Sophie Chattington at the University of Bremen in Germany, while the other two partners are at IVIA in Valencia, Spain: *Amblyseius swirskii* along with Angeliki Paspali, and *Nesidiocoris tenuis* with Milena Chinchilla-Ramirez. It's this final species that I will talk about today.





Meet Nesi

- *Nesidiocoris tenuis*, mirid
- World-wide distribution
- Used throughout Spain against tomato pests
 - *Tuta absoluta*, whitefly, spider mites, thrips
- Zoophytophagous

Aspects that complicate sequencing

- Diploid
- Generation time: ~1 month
- No reference genome
- Unknown :
 - Inbreeding ability
 - Standing genetic variation

With that, I'd like you to meet *Nesidiocoris tenuis*, which we shorten to Nesi. Belonging to the family Miridae, it has a worldwide distribution, but is used as a biocontrol agent nearly exclusively in Spain. It is used in tomato crop against *Tuta absoluta* and whitefly in particular. Why only Spain? One reason is that, like most mirids, it is zoophytophagous; when Nesi runs out of pest to prey on, it will start eating the plant. The arrows in the image to the left are all signs of crop damage, such as necrotic rings, wilting leaves, and even fruit loss. Yet, Spanish farmers swear by it. So there's room for improvement. That's what Milena's project tackles, whereas I'm more interested in the genome. There are key aspects to Nesi that complicate sequencing: it's diploid, with fairly long generation time of 1 month for egg to egg-bearing adult. It also has no reference genome (while a mitochondrial genome has been published). Additionally, it's unknown if it's able to be inbred, and the standing genetic variation of the commercial population is a mystery.

Goldilocks & the 3 Sequencing Strategies

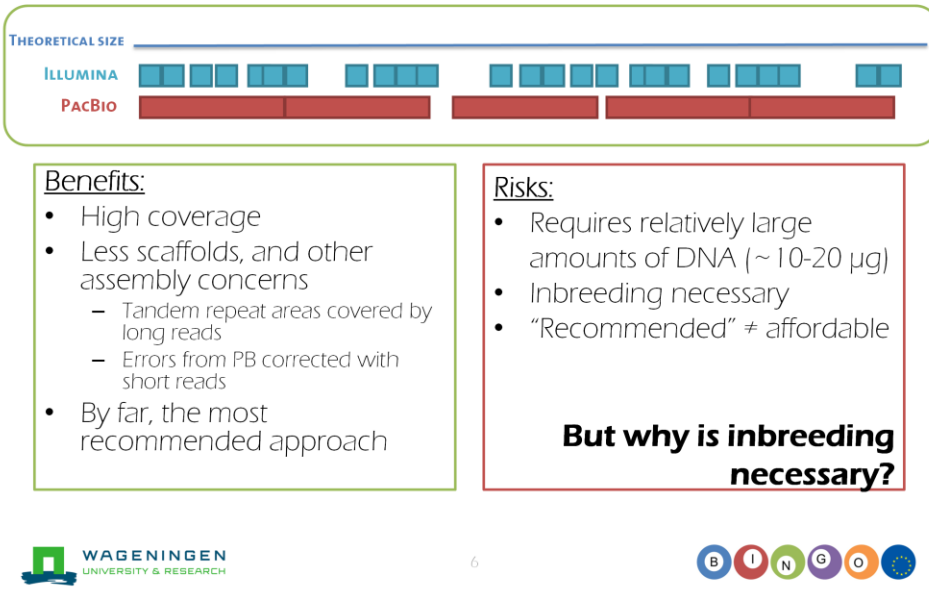


Arthur Mee and Holland Thompson, eds. *The Book of Knowledge* (New York, NY: The Grolier Society, 1912)



I've been in this project for about a year and a half, and it became apparent that I was searching for the perfect sequencing strategy for Nesi. In fact, you could call this story Goldilocks & the 3 Sequencing Strategies – I'm Goldilocks. If you're unfamiliar with the British fable, Goldilocks is a little girl walking through the woods. She comes across the 3 Bears' house and (essentially) breaks in and faces choices. For instance, the bed are either too hard, too soft, or just right. The porridge is too hot, too cold, or just right. This search for the perfect middle point between extremes is used in other fields of research, such as astronomical research into faraway planets that could support life: they exist within the Goldilock's zone. It's with that search for the perfect balance that I waded into my findings.

1st Strategy: PacBio/Illumina hybrid *de novo*



After asking a few experts in insect genomics, I decided that a hybrid *de novo* approach would be the best option to pursue, when looking at the ease of assembly, as well as the amount of coverage possible for those tricky tandem repeat areas. Here you can see the benefits and the risks. But why is inbreeding necessary?

Inbreeding for homozygosity

- Why? As little polymorphism as possible
 - $2n \rightarrow n$
 - Error rate of PacBio and ease of assembly
 - Decrease complexity \rightarrow genomes are a generalization



3 attempts at lab rearing over 1 year
None survived to the 2nd generation

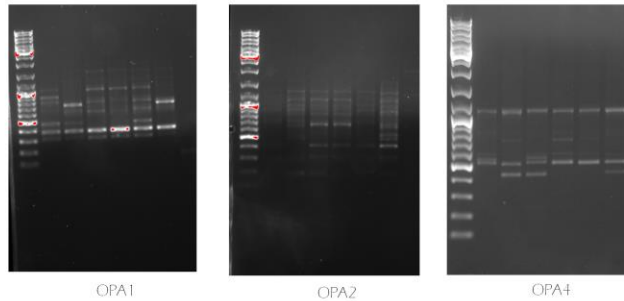
- 1 month/generation \rightarrow 10 inbred generations = 10 months
 - not sustainable to keep trying past 1 year
- Inbreeding may be possible, but impractical for this project

Maybe it's already homozygous? Or close?

There are a few reasons, but it all leads to reducing the complexity of the final genome and easing the assembly stage. Heterozygous genomes would require phasing, and if there are several heterozygous individuals in a sample for sequencing, the resolution of the resulting genome would be tricky: What is an error, what is a SNP? When does a tandem repeat area end? This also comes into play with the PacBio part of the hybrid strategy, which can be error prone. All of these are good reasons to inbreed. However, all three of my attempts to rear were foiled – I could not get the Nesi to a 2nd generation. And if it takes 1 month per generation, I would need 10 months to achieve 10 inbred generations (the amount that is often quoted as being the bare minimum). 10 months is a long time considering this is step 0 of the genome assembly. So while inbreeding may be possible, it is impractical for my project. But, what if the commercial strain is already (near) homozygous?

Well, how variable is the population?

- Single-source population for more than a decade
- Maybe there's little variation i.e. inbreeding not req'd
- Random amplified polymorphic DNA (RAPD) markers
 - 6 individuals, 10 RAPD markers (OPA series 1-10)



OPA1

OPA2

OPA4

Ferguson, KB (unpublished data)

What if the Nesi are already inbred? This idea goes back to the history of the commercial strain. They were collected from a single population in relatively small numbers, and have existed without genetic refreshment for quite some time. So maybe a mix of initial bottleneck and genetic drift have led to reduced population variation. I tested this with 10 RAPD markers, hoping that six individuals would show significantly similar banding patterns. Well, this wasn't the case. The three PCR images here show the ladders on the left and the individual samples. They are quite different, suggesting that some variation must still exist. And while RAPDs are not the most reliable method, it was enough to discard this hope that the population had reached low variation.

1st Strategy: PacBio/Illumina hybrid *de novo*

THEORETICAL SIZE

ILLUMINA

PACBIO



Benefits:

- Increased coverage
- Less scaffolds, and other assembly concerns
 - Tandem repeat areas covered by long reads
 - Errors from PB corrected with short reads
- By far, the most recommended approach

Risks:

- Requires relatively large amounts of DNA (~ 10-20 µg)
- Inbreeding necessary
- “Recommended” ≠ affordable

**Without homozygosity,
too risky to pursue**

Without homozygosity, the hybrid *de novo* is too risky to pursue. So I moved on to the next bowl of porridge.

2nd Strategy: i5K approach



- Specifically, their protocol for difficult to inbreed species
- Pooled individuals for a larger sized library , and single individual for smaller insert sizes

Benefits:

- Homozygosity not required
- Established assembly pipeline

Risks:

- More time-consuming assembly method
- Enough material from individual

There is a Dutch saying: Better a good copy than a bad original. I took this to heart and looked to the i5K project for guidance. If you're not familiar with the project, it is a worldwide initiative to sequence 5000 insect genomes, with an initial 28 being wrapped up soon. They have a protocol in case of species that are difficult to inbreed. It involves getting enough DNA from a single individual as well as pooled samples. The benefits and risks are listed here.

2nd Strategy: i5K approach



- Specifically, their protocol for difficult to inbreed species:
- Pooled individuals for a larger sized library , and single individual for smaller insert sizes

Benefits:

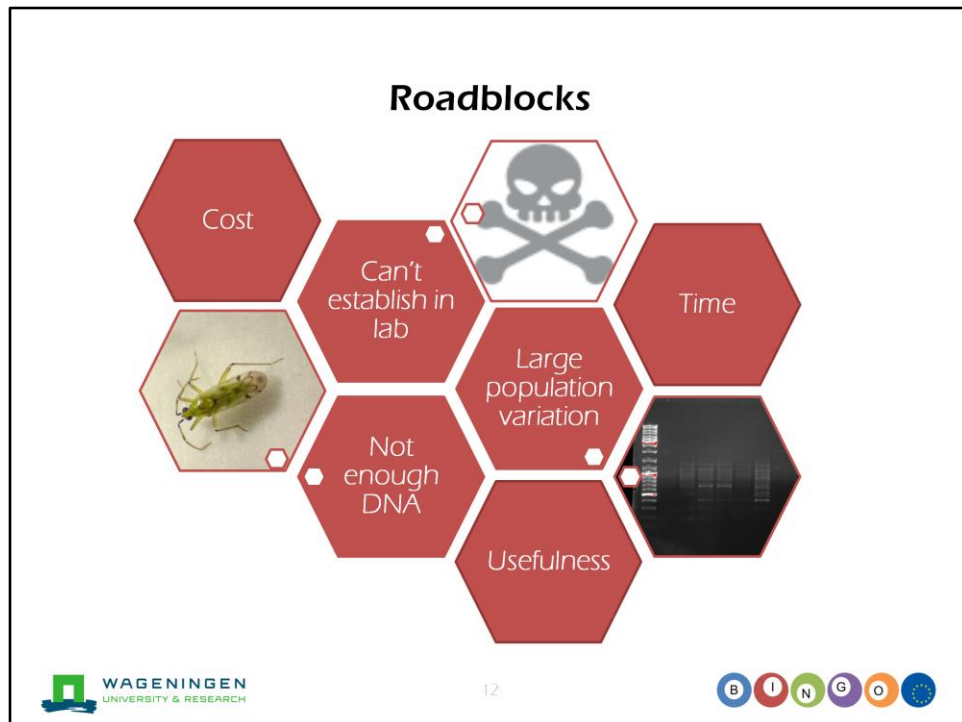
- Homozygosity not required
- Established assembly pipeline

Risks:

- More time-consuming assembly method
- Enough material from individual

Not enough DNA from an individual Nesi

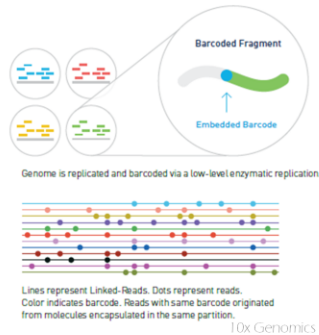
Unfortunately, it may have been too soon to see that as the alternative, as I cannot get enough DNA from Nesi! This strategy was used in mountain pine beetle to success, but it seems that Nesi is too small to get enough high quality DNA for this approach.



And so I faced these roadblocks: the inability to establish in my lab, the seemingly large existing population variation, and that I can't get enough DNA from an individual. On top of this, I have to consider cost of sequencing strategy, the time it will take to rear as well as assemble, and finally the usefulness of the genome. We could have done some shallow population-level sequencing, but that doesn't help the argument of biocontrol using genome research for improvement, nor does it help my collaborators. It was at this point I became demotivate with this search. Nesi became my least favourite child – I focused on the other species, know that I on the backburner I had this seemingly impossible project. Thankfully, a third bowl of porridge arrived!

3rd Strategy: 10x Genomics

- The 10x Chromium library prep uses barcodes on short sequences
- Illumina for the sequencing



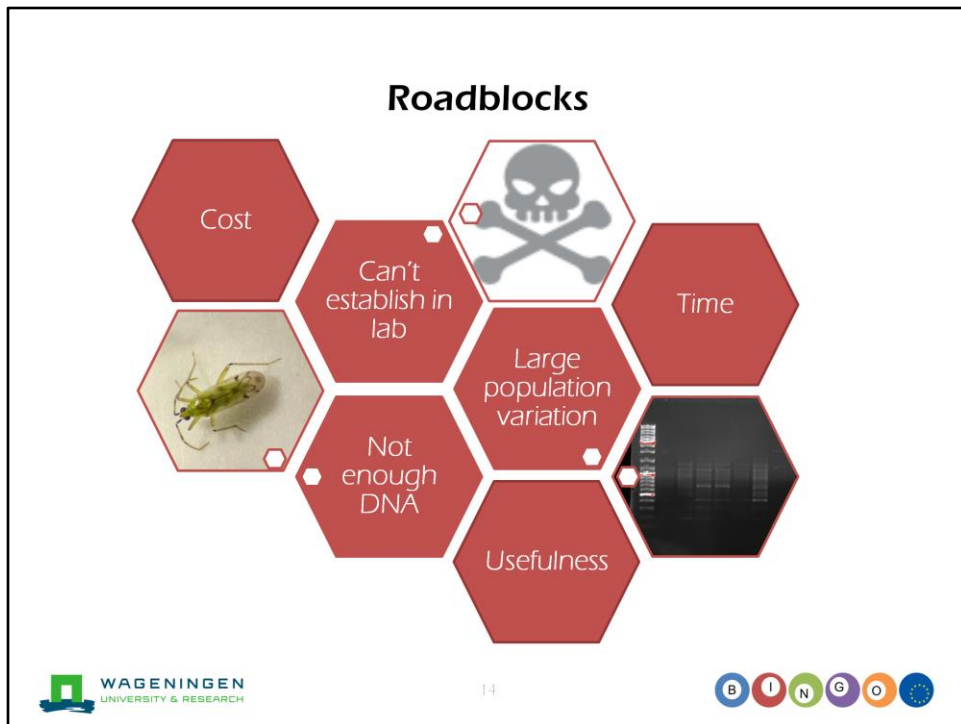
Benefits:

- High coverage
- Less scaffolds
- Small amount of DNA
- Phased genomes with structural variants

Risks:

- Proprietary library prep, assembly algorithm, and visualizations
- Cost prohibitive
- Chromium reaction could fail

My supervisor, Bart Pannebakker, sent me to our Wageningen colleague Elio Schijlen to talk about 10x Genomics. 10X Genomics approach is essentially a specialized library prep – the sequencing occurs on Illumina. And to be honest, this whole process is still a bit like snake oil – I’m told it works and how but it’s still a bit fuzzy. So I encourage you to find more brilliant minds to break it down, but I will attempt: Essentially, small amounts of DNA from a single individual are separated by microfluidics into small droplets with mineral oil. Molecular barcodes specific to the small droplets attach to the DNA and are incorporated into the sequence. After the Illumina sequencing step, the barcodes are used to assemble the genome, resulting in incredibly high coverage. You can find more information from 10x Genomics. There are risks associated with the 10x Genomics approach, as described on the slide, but the benefits are huge: high coverage less scaffolds, as well as phased genomes with structural variants. And small amounts of DNA are seen as a positive.



This means that some of my roadblocks, the inability to establish as well as the small amount of DNA from an individual, go from being problems...



Jessie Willcox Smith - Goldilocks and the Three Bears, 1916

The End?

So far:

- BioScience at Wageningen University
- Confirmed 10x barcoding reaction
- Currently being sequenced
- Bioinformatics to be performed soon
- Goal: Collaborative, open annotation strategy

Is this the end for me, Goldilocks? Well, for now, I'm waiting to hear back from my partners in Wageningen who are doing the library prep, sequencing, and assembly. Currently, I have been told that the barcoding reaction worked and am waiting for it to be sequenced and assembled. Because of the proprietary nature of the 10x Genomics method, I will have more time to focus on the annotation, and aim to have an open, collaborative annotation strategy, so that should be coming up soon.

Acknowledgements

- My supervisors:
 - Dr. Bart Pannebaker, Dr. Eveline Verhulst, and Dr. Bas Zwaan
- Laboratory of Genetics, Wageningen University
- Dr. Elio Schijlen at BioScience (Wageningen University and Research)
- Koppert Biocontrol Systems
- Milena Chinchilla-Ramirez and IVIA



17

I'd like to acknowledge the following people, specially my supervisors, my BINGO colleagues, and those at BioScience here in Wageningen, Koppert Biocontrol Systems for providing the Nesi, and IVIA for some images here. For more information on the BINGO project, go to www.bingo-itn.eu, where you can find out all about the network and the various projects within it.



Questions?

Kim Ferguson



18

You can follow me on Twitter, @kfergy. I am also on ResearchGate and LinkedIn.

Photo Credits

- Slide 2: BINGO-ITN (www.bingo-itn.eu)
- Slide 3:
 - *Trichogramma brassicae*: ©BugsInThePicture
 - *Amblyseius swirskii*: ©Koppert BV
 - *Nesidiocoris tenuis*: Kim Ferguson, 2017 CC by 4.0
- Slide 4:
 - Tomatoes: Kim Ferguson, 2017 CC by 4.0
 - Tomato plant: Instituto Valenciano de Investigaciones Agrarias (IVIA)
- Slide 5: Arthur Mee and Holland Thompson, 1912
- Slide 10: i5K logo, <https://www.hgsc.bcm.edu/arthropods/i5k>
- Slide 13: 10x Genomics promotional image, 10genomics.com
- Slide 16: Jessie Willcox Smith, 1916
- Emojis are provided by EmojiOne (2.3)

Note that the images on slides 5 and 16 are in the US Public Domain.