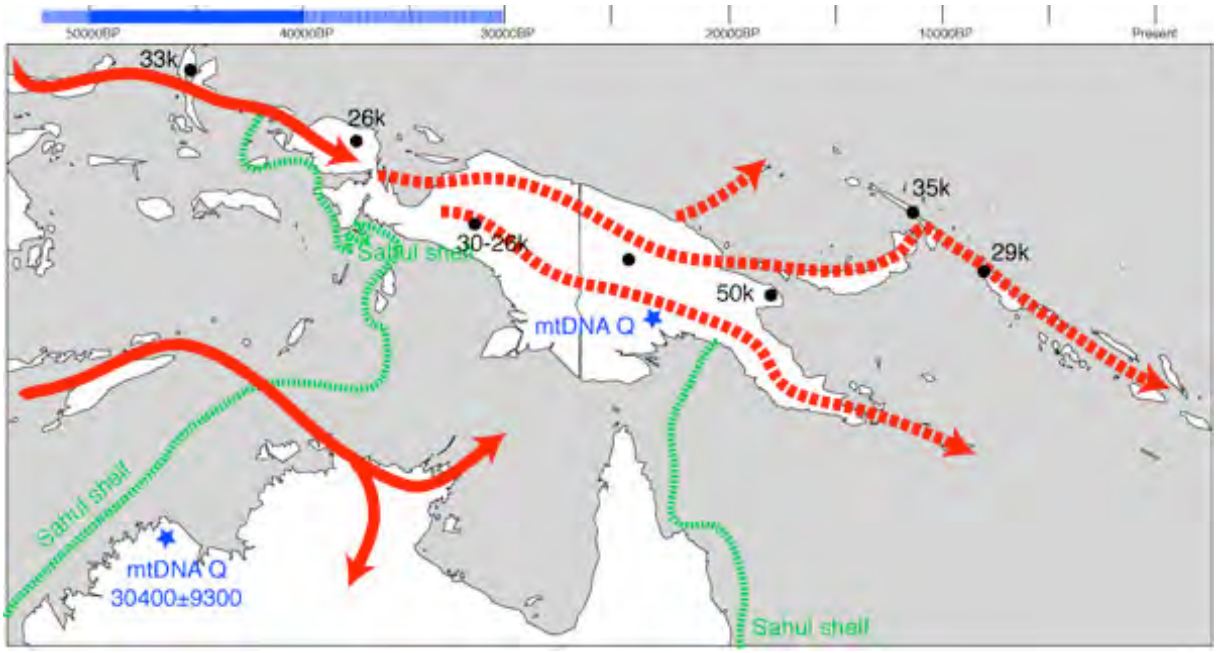
Linguistic traces of the colonization of New Guinea

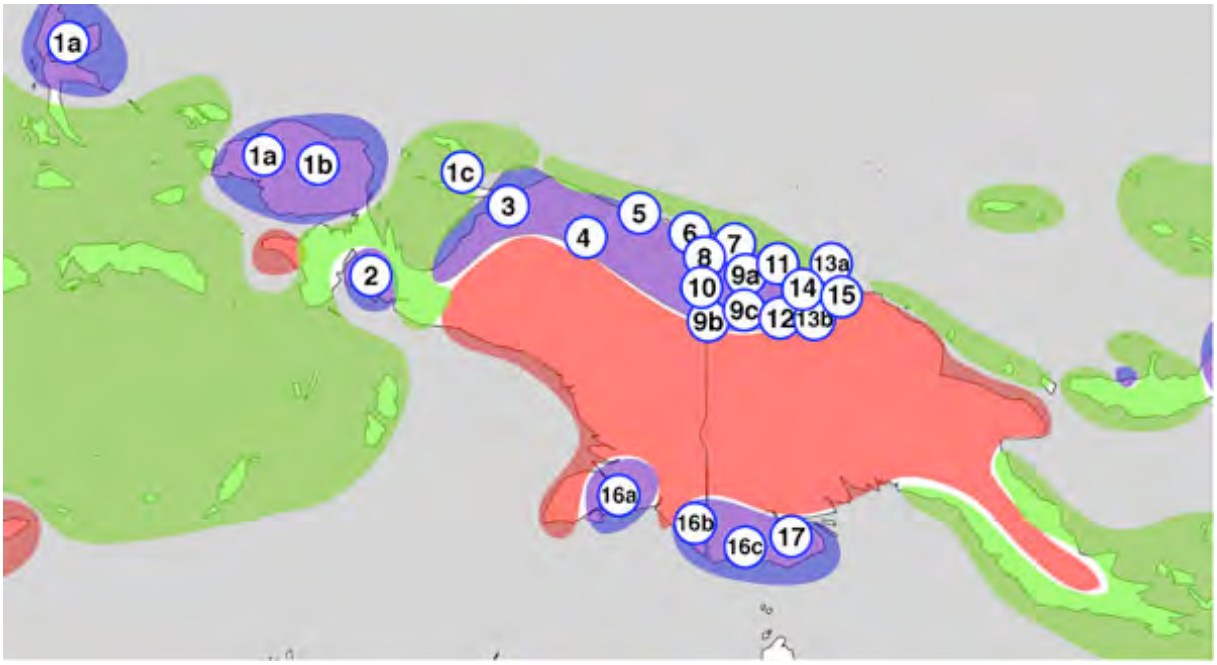
Michael Dunn & Ger Reesink
Max Planck Institute for Psycholinguistics
Radboud University, Nijmegen

How can linguistics contribute to our knowledge about human dispersals in the distant past? We will consider the case of New Guinea and surrounding islands, one of the most linguistically diverse areas of the world. This study is a follow-up on the Eurocores OMLL project *Pioneers of Island Melanesia*, reported in Dunn et al. (2005).



Earliest migrations: two or more?

A possible scenario would assume at least two major migration (Summerhayes 2007, see above) waves through Wallacea into Sahul, perhaps the oldest one, ~40,000 BP, following the northern route (Sulawesi, Halmahera, Bird's Head and further to the east along the north coast), the ancestors of non-TNG, and a second one, ~20,000 BP, through the Lesser Sundas directly onto present-day north Australia and Aru island, with a northward trek into the Highlands, the ancestors of TNG. This scenario would have the TAP and, possibly, the South Papuan families as stay-behind descendants of the TNG precursors.

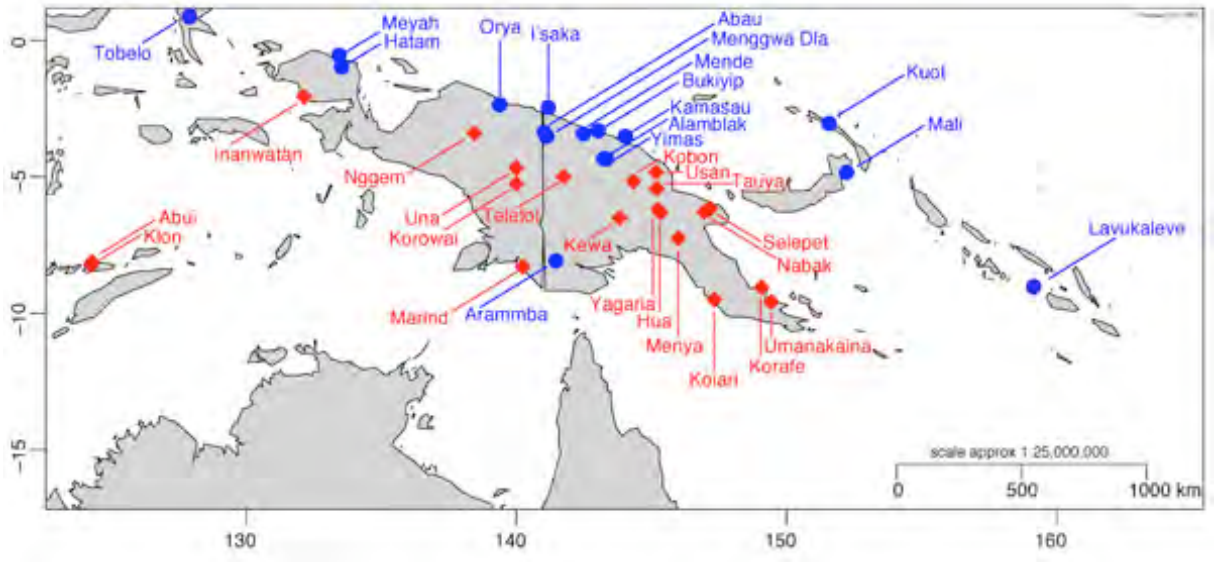


Papuan Language Groups (Ross 2005 Proposal)

Recently, Ross (2005) has proposed the most promising classification on the basis of pronoun sets of more than 600 Papuan languages. He explicitly offers his classification as

> "part of a tentative application of Step 1 of the Comparative method: 'Determine on the strength of diagnostic evidence that a set of languages are genetically related'; that is, I am offering diagnostic evidence of a family, and weak evidence at that" (Ross 2005:49).

His proposal reduces the number of Papuan families to 23, including the large TNG family with many daughter families and isolates, and nine isolates. Our aim here is to investigate to what extent the distribution of abstract structural features captures this classification.



Map with TNG & non-TNG samples

What does it mean to say that two entities, in this case languages, are related? In population genetics care is taken to sample DNA from unrelated individuals, but subsequent

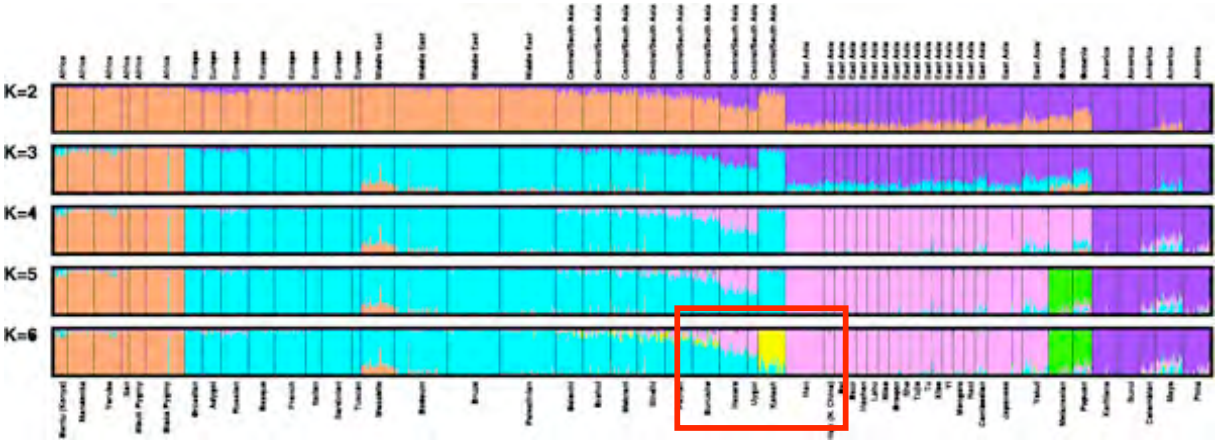
analysis of their mtDNA, Y chromosome or alleles of their autosomal markers, will classify them in genetically related lineages.

The comparative method is able to identify language families through cognates in lexical items and morphological formatives exhibiting regular sound changes. But this level of relationship has limited time-depth, unlikely to exceed 10,000 years. The comparative method requires a significant number of cognates, or a well-defined regular sound change in the case few cognates are left, or a small paradigm of irregular forms to rule out chance resemblances. That is, a single or few resemblant lexical items may still suggest genealogy when the probability of the number of shared phonemes is so small that it can be considered individual-identifying (Nichols 1996).

When none of these conditions are met, languages are said to be unrelated. But when this involves languages in close propinquity that share a sizeable number of structural features, the interesting question is: how did those languages get to be near neighbors, both geographically and structurally? Did they diverge a long time ago and could they still be related as far relatives or did one one of them immigrate from far away?

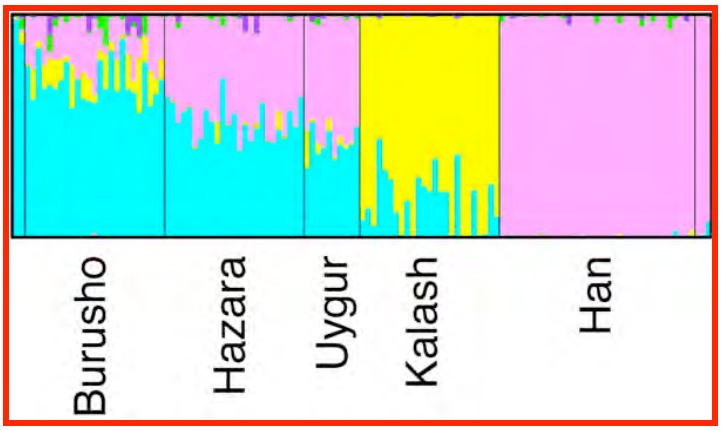
On the basis of a profile of 170 binary structural features, we are investigating historical relationships between the proposed Papuan groups.

The selection of languages for this study was based on the preliminary classification of Papuan languages on the pronoun sets in Ross (2005). We set out to investigate a sample of 5% of the roughly 800 Papuan languages, representative of both the known or suspected subgroups and of their geographical distribution. So far we have coded 20 languages of the proposed TNG family and 16 belonging to non-TNG groups.



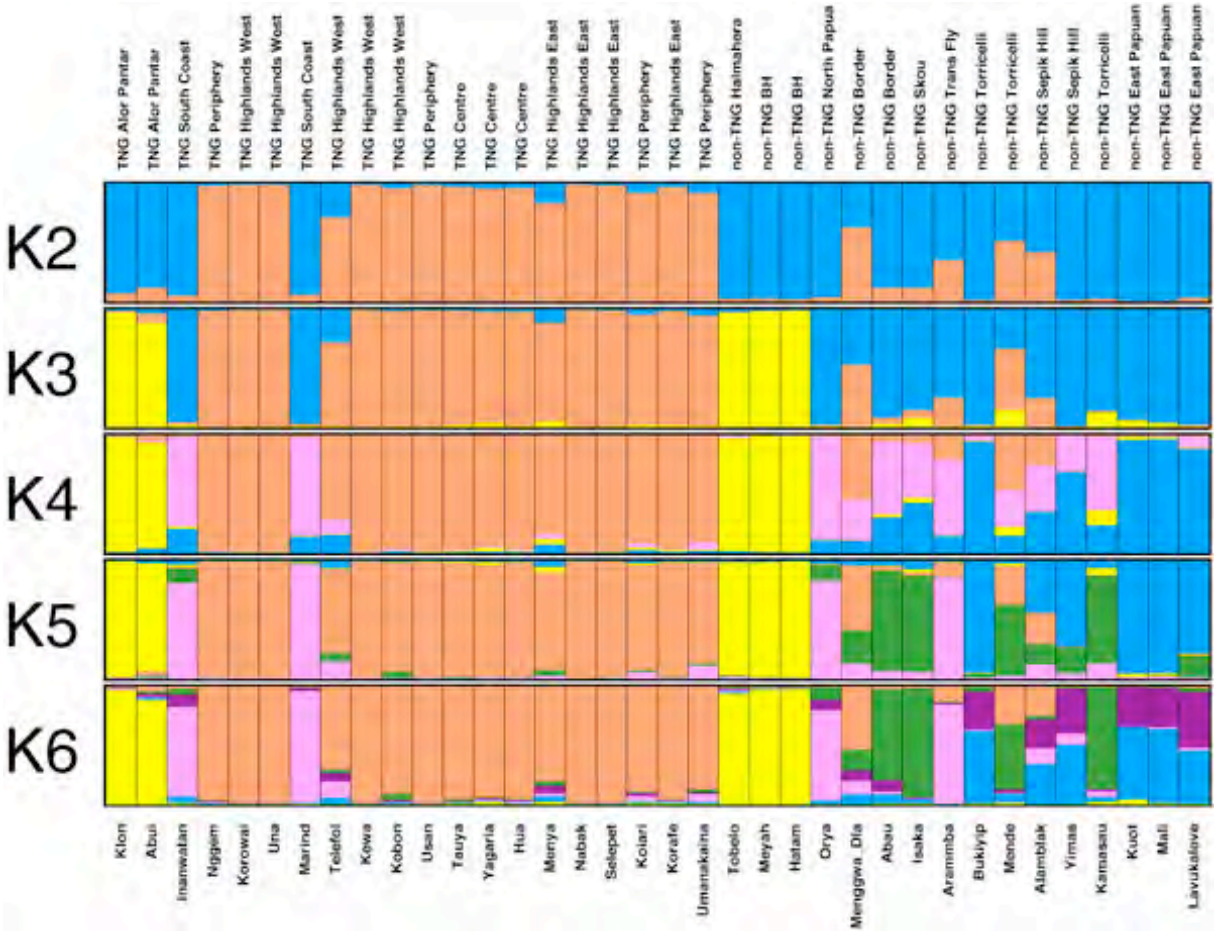
'Structure'

Humans do not just inherit mtDNA in the maternal line and Y chromosome in the paternal line of descendancy, their individual genetic markers are formed by recombining markers from both parents, all four grandparents, and so on. In human genetics a clustering algorithm has been developed, called structure (Rosenberg et al. 2002. Genetic Structure of Human populations. *Science* Vol.298:2381-2385), which identifies subgroups that have distinctive allele frequencies. Rather than starting with predefined "populations" and investigating how they differ in terms of presence or frequencies of certain mutations, this program places individuals in K clusters. The value of K is chosen in advance, but can be varied across independent runs of the algorithm. Individuals can have membership in multiple clusters, with membership coefficients summing to 1 across clusters. Such multiple membership would indicate different contributions from ancestral populations showing more or less admixture.



'Structure' detail: Kalash as isolated population in central Asia

This method identifies at K6 an isolated group in northwest Pakistan, speaking an Indo-European language



'Structure' analysis of TNG & non-TNG

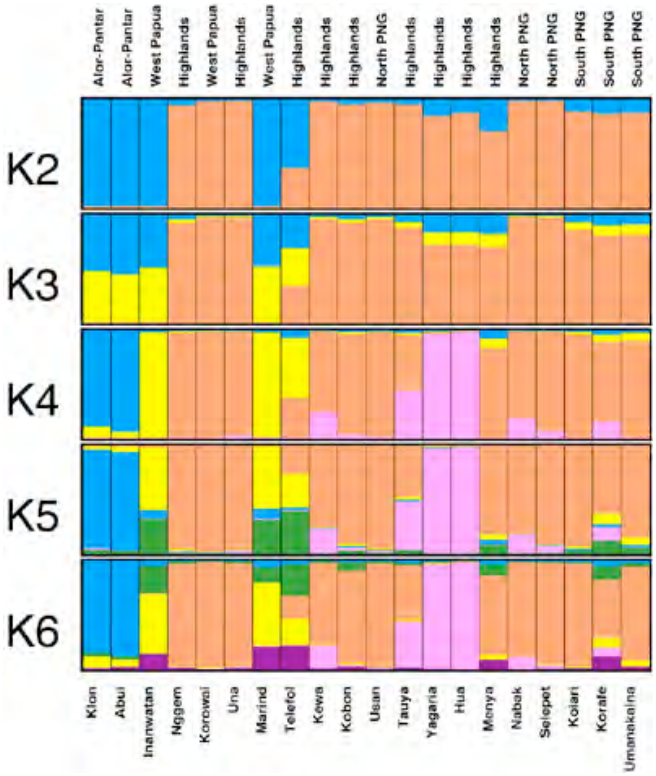
The computational instrument STRUCTURE treats each of the 36 Papuan languages in our sample as an individual which has recombined features inherited in different proportions from multiple ancestors. A few independent runs showed a high degree of congruence in the log likelihood scores for values K2, 3, 4 and 7, with K3 yielding the highest probability.

This slide shows the inferred population structure for each language in independent runs identifying K2-10. The columns show individual languages, labelled by name below and by rough affiliation and geographic location above. At K2 there appears a rather clear

dichotomy between TNG and Non-TNG languages, maintained for all K values. It is obvious that the TNG languages remain as a solid block, while finer levels of granularity exhibit a greater diversity of the Non-TNG languages.

Significant exceptions in the TNG sample are four languages in the west and southwest extremes of the TNG expanse. At K2 these languages, Klon and Abui of Alor, and Inanwatan and Marind of the South coast of Indonesian Papua, are grouped very clearly with the Non-TNG population, in contrast to some Non-TNG, Menggwa Dla, Mende, Alamblak of the Sepik basin and Arammba of the Trans-Fly, that show contributions from both.

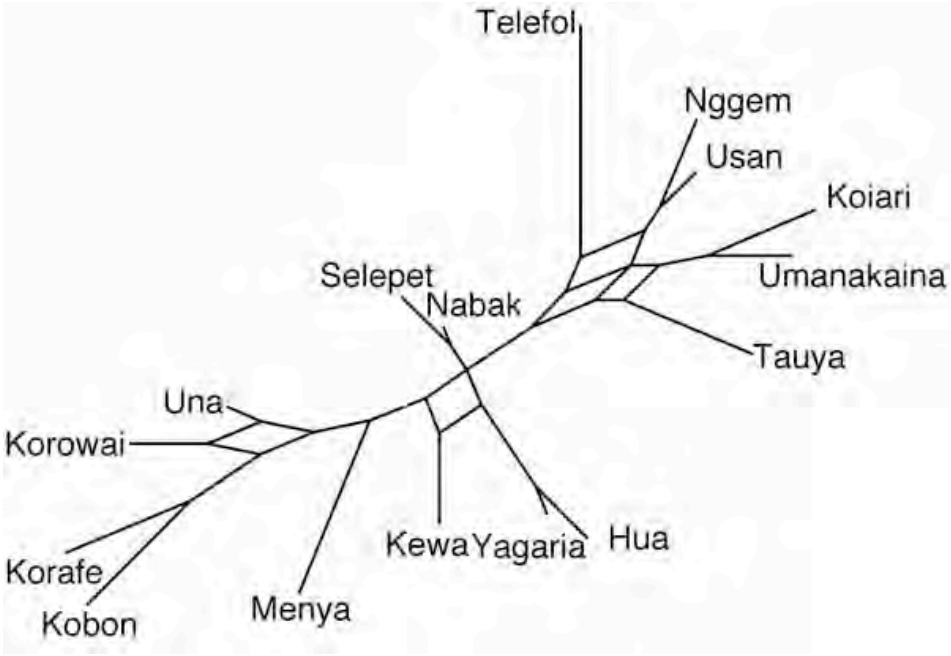
Interestingly, when the program is told to search for three contributing populations, Inanwatan and Marind remain with the bulk of the Non-TNG, while the Alor languages are grouped with Non-TNG from Halmahera (Tobelo) and the Bird's Head (Hatam and Meyah), as a West Papuan population..



'Structure' of TNG

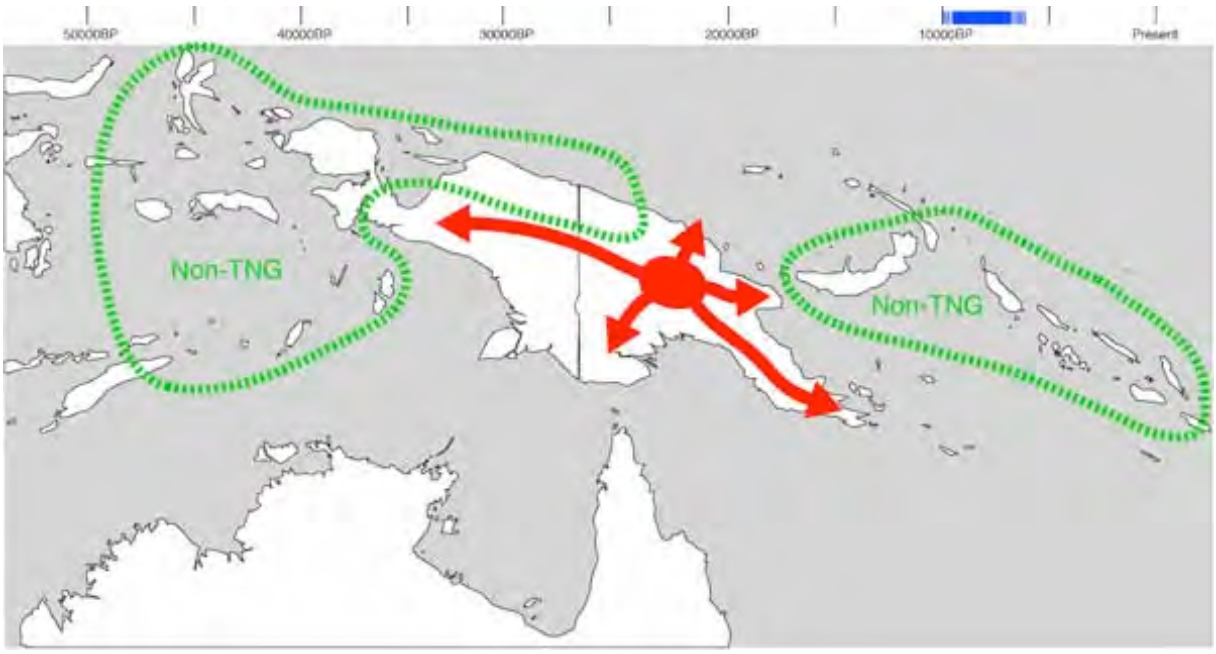
Since the STRUCTURE analysis of all Papuan languages does not differentiate between the languages of the TNG family, it pays to subject this group of languages to a separate analysis. We will include the putative members, Klon and Abui of Alor, and Inanwatan and Marind of the Papua South Coast.

A first hypothesis based on this analysis would posit five to six populations, not necessarily in chronological order, as branches from Proto-TNG, assuming its homeland is between the Strickland River and the Eastern Highlands province, with the center of gravity in the eastern Highlands, Chimbu and western Highlands provinces of present-day Papua New Guinea.



Consensus Bayesian Network of TNG

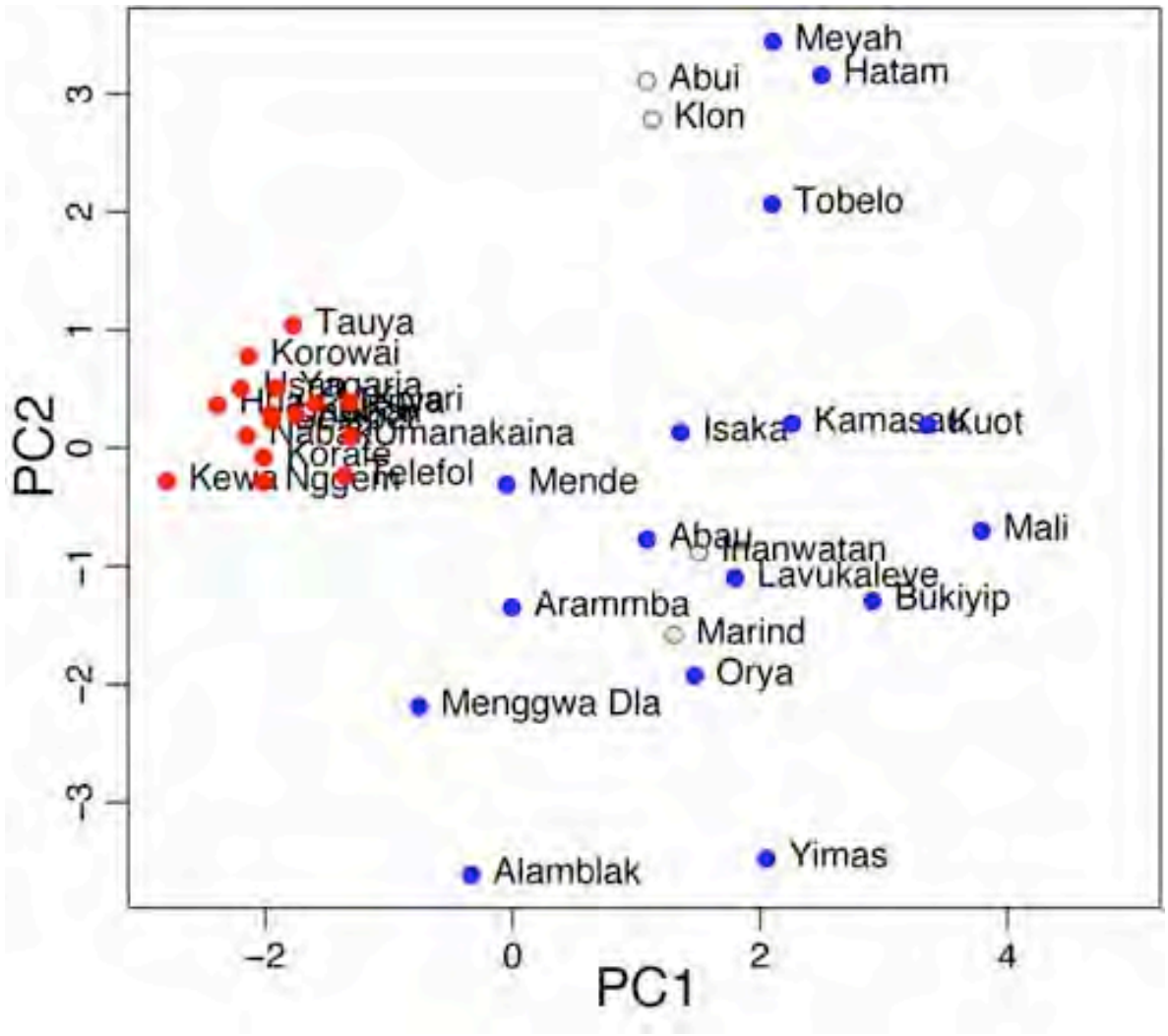
Since we have reasons to doubt the membership of the outliers of the TAP and Papua South Coast families, we applied the Bayesian phylogenetic inference on the what we believe is the best sample of the TNG family. We will return to the position of the outliers in the proposed migration patterns by which New Guinea has been colonized in the conclusion.



Map showing star-like expansion of TNG

The network obtained by the Bayesian inference suggests a star-like expansion of the TNG family, with a cluster containing languages on the peripheries:

1.Nggem and Telefol in the west, Usan and Tauya in the north, and Koiari and Umakaina in the southeast.
2. Una and Korowai in the west; Kobon in the center; Korafe in the east.
3. Stay-behind in the center: Hua, Yagaria, Kewa, and Menya, with an off-shoot to the northeast: Selepet and Nabak.

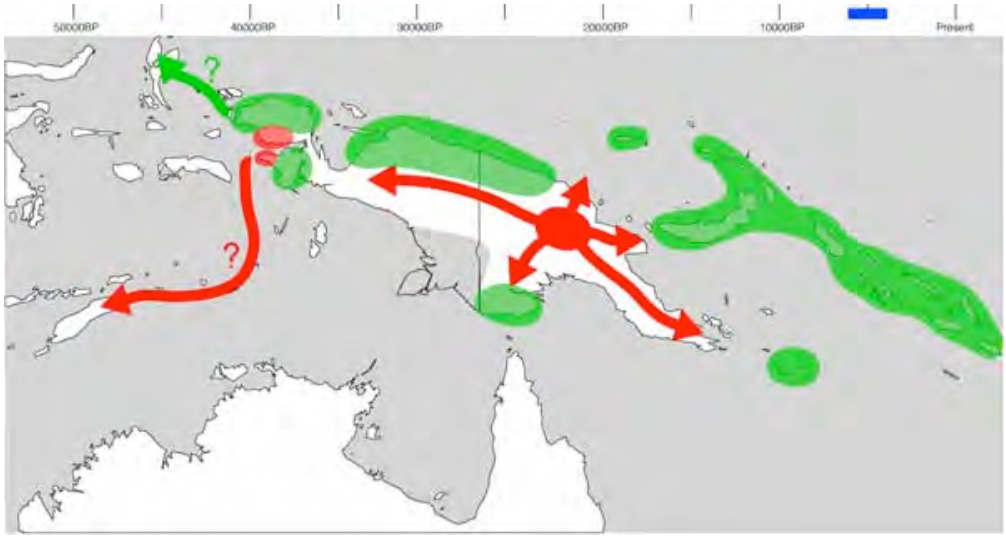


Pc1 x PC2: TNG versus non-TNG and West Papuan

A Principal Component analysis yields two clusters of features that contribute significantly to the observed diversity. For PC1 the main criteria appear to be a positive loading of noun classes and gender marking, subject prefixation on the verb and VO order with concomittant prepositions. Negative loading on this dimension is basically Tense marking on the verb, suffixing of subject and OV order, postpositions, and clause chaining with a switch reference system.

This PC basically divides the TNG languages and the Non-TNG languages. To any Papuanist this result comes as no surprise, see Wurm (1975), Foley (1998, 2000), Reesink (2005). At the same time, this result provides statistical robustness to the claim, rather than just an impressionistic characterization of both Papuan groupings.

PC2 includes a few phonological features, both positively and negatively. In the nominal domain, the positive loading involves alienable versus inalienable possession, against a negative loading of Noun classes and genders. In the verbal domain, this dimension is defined negatively by (absence of) Tense marking and suffixing of subject (A,S). As the scatterplot shows, PC2 makes a strong division between the West Papuan languages and a number of languages of the Sepik basin, with the rest clustering somewhere in the middle.



Map: Back migration before AN influx; Sepik-Ramu; TNG

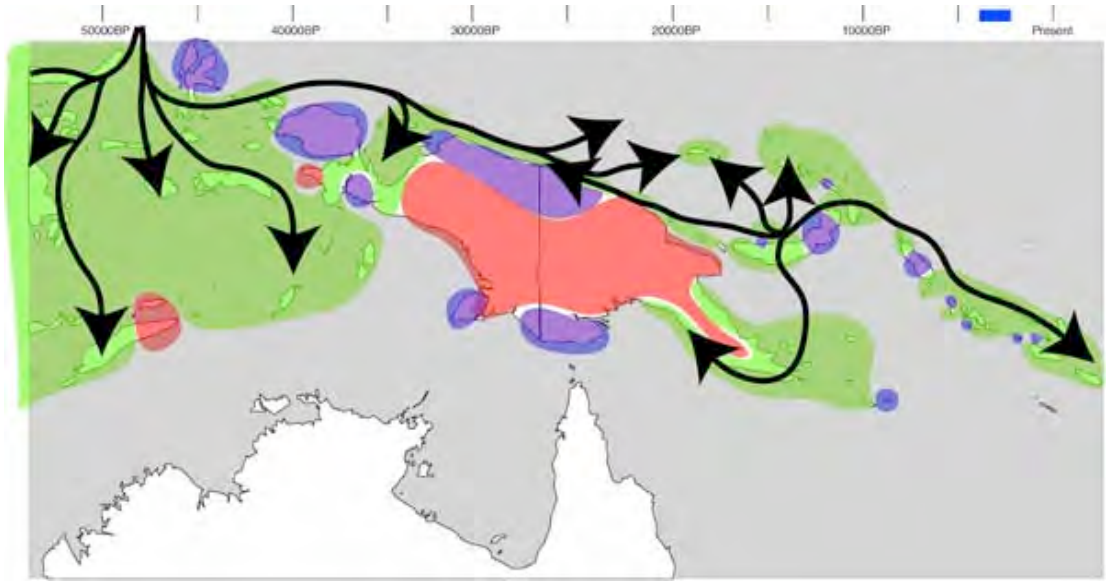
It seems plausible that the oldest stratum, representing the first colonizers of New Guinea, diversified in the 23 and some families plus isolates, arrived at by Ross (2005), which includes the precursor of TNG. It was only in the early Holocene (Ross 2005:41; Pawley 2007:50) that the emergence of agriculture in the eastern Highlands give rise to a population expansion and subsequent migrations of the various branches of TNG.

As regards this scenario, Wurm et al. (1975) said of one of their proposed early Papuan populations in New Guinea "It seems tempting to suggest that this far-flung substratum which may perhaps have surviving primary manifestations in some members of the West Papuan Phylum [. . .], in the Torricelli Phylum and perhaps also in the East Papuan Phylum [Island Melanesia], may outline the earlier presence in the New Guinea area, of an old language type [. . .] to be later overrun and reduced to substratum level by subsequent language migrations". Wurm et al. characterized the languages in question as having a prevalence of certain pronoun forms, an overt two-gender system, a tendency to prefixing, number marking on nouns, verb supplementation and alternation in connection with object and subject marking, and the absence of medial verb forms (pp 940–941).

As we have seen, these features are part of the cluster that defines the PC1 which separates TNG and non-TNG languages, and they corroborate at least the genealogical unity of the TNG family. The relationships within the much more diverse non-TNG groups are much less evident. This may be due to the fact that the TNG has a lesser time-depth than its earliest relatives, assuming a singular colonization event, as suggested by the human genetic data (Hudjashov et al. 2007).

This scenario leaves us with the enigma of the putative TNG languages, still spoken today on Timor, Alor and Pantar and the South Papuan languages Marind and Inanwatan. Their membership of the TNG is based on rather slender evidence (Pawley 2007:47). Yet, it is claimed that they are the result of a east-west migration of a TNG branch, preceding the arrival of the AN languages in the Moluccas (Ross 2005:42).

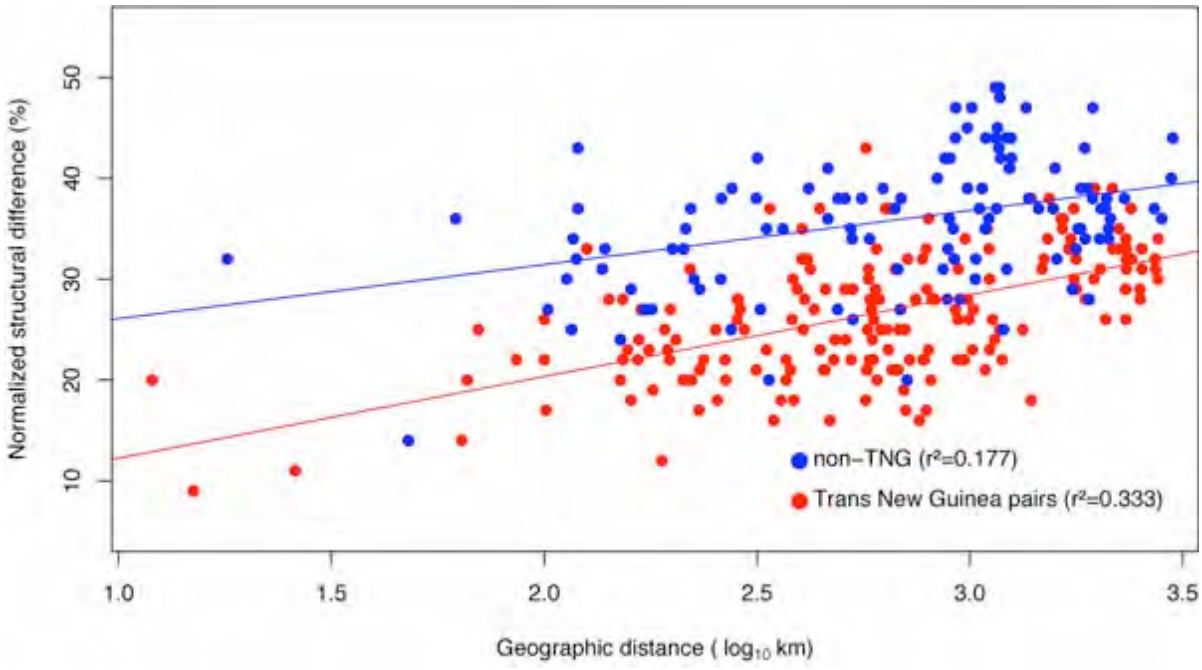
Leaving the complex situation in the western Papuan region aside, we note that the expansion of TNG apparently has driven two wedges between non-TNG speaking populations: (i) between languages along the north coast and the remnants found in the Trans-Fly area; and (ii) between languages along the northern rim of New Guinea: there are no non-TNG languages between the Lower Sepik and Lower Ramu and Island Melanesia. All of Madang and the Finisterre-Huon are clearly members of the TNG. As Pawley (2007:50) notes, "The high degree of diversity within the Madang subgroup points to a very early TNG presence in the eastern half of Madang province".



Austronesian expansion into Moluccas and along north coast of NG into Island Melanesia

The Austronesian family is the largest family inthe world, in terms of numbers (about 1,200 languages), as well as in terms of geographical extension: from Madagascar to Easter Island, from Taiwan to New Zealand. Generally, the history of the various migrations and the resulting subgroups within this family are well-known. The AN homeland is Taiwan (possibly south China). Starting about 4,000 years ago, various migrations spread from there south through the Philippines, present-day Indonesia (from there westward to Madagascar), the Moluccas and the Cenderawasih Bay. From there the ancestral language(s) of the Oceanic subgroup moved further east (approximately 3,200 years ago) to their homeland on New Britain, from where the Oceanic speakers moved further east, south, and back to the west.

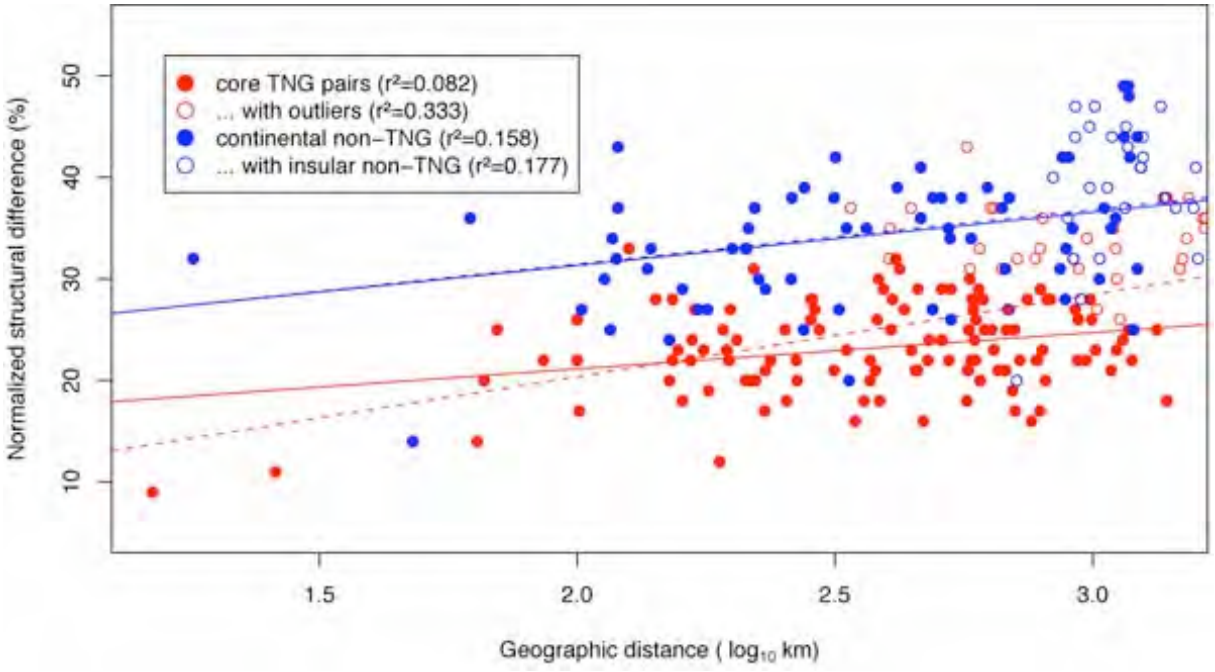
Not only in the Moluccas and along the NG north coast, but also throughout Island Melanesia, there is ample evidence of both linguistic and genetic admixture between AN and Papuan speaking groups (e.g. Friedlaender (ed) 2007).



Spatial-Structural Correlation

Since linguistic transmission is both vertical and horizontal, languages that are geographically close may resemble each other either because of shared ancestry or because of contact diffusion or both. In order to test whether resembling profiles for presence and absence of strucural features between languages are the result of just areal contact (this is the criticism from comparative linguists leveled against the use of typological features), we calculated the structural and geographical distance between each pair of languages in our sample. Structural difference is the percentage of features on which two languages score a different value, out of all the features which had an actual score 1 or 0. Geographic distance is based on the latitude and longitude coordinates for roughly the center of the language area.

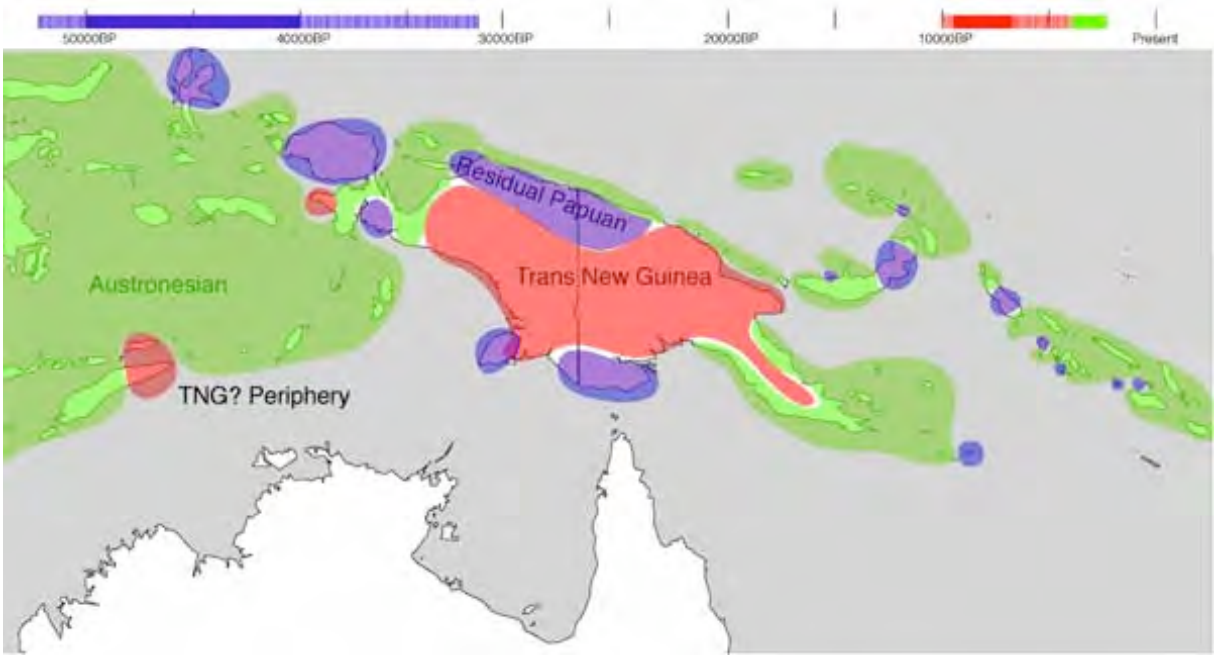
It can be seen that indeed geographic distance correlates with structural distance. This correlation is about twice as strong for the TNG languages (R^2 = 0.333) than it is for Non-TNG languages (R^2 = 0.177). Thus, some areal signal seems to be present, but this can still be due to two factors, genealogy and contact, or both.



Geographic outliers

Since extreme geographic distances, either very close or very far apart, may actually enhance the correlation between structural ~ spatial distances, and because the STRUCTURE analysis suggests that the putative outliers of TNG may not belong to that family, we did the correlations for TNG twice: 1) without the Alor languages Klon and Abui and the South Papua languages Inanwatan and Marind, and 2) with these outliers included. We also calculated the correlation for the non-TNG sample twice, one including the insular outliers (East Papuan and Tobelo) and one without them.

The results show that for the pruned TNG languages, the spatial correlation is virtually absent, while for the Non-TNG languages there is hardly any difference. This finding is commensurate with the hypothesized star-like expansion pattern, based on STRUCTURE and the Bayesian phylogenetic inference tree. Thus, it would be difficult to contribute the distribution of cluster of features that characterizes this family to only areal diffusion.



Summary todate and remaining questions

We have discussed possible migrations that account for the distribution of the major linguistic groupings found in the larger New Guinea area

Desiderata

There are a few questions remaining before we can address these speculations with more confidence:

(i) A much wider sampling of Non-TNG languages is needed, to ascertain whether they form a unit as solid as the TNG.

(ii) Comparison with east Indonesian AN languages is necessary, to investigate further genealogical or contact-induced clusters; similarly a comparison with Island Melansian oceanic is called for.

(iii) Further research is necessary to determine dependencies (logical, typological, correlated evolution) of various features and how that would influence our analyses.

(iv) We need to formulate more precise hypotheses about past migrations and how they can be falsified.

(v) Given the phylogenetic signal carried by a cluster of features, it may be that lexical resemblances can strengthen relationships.

References

DUNN, Michael, TERRILL, Angela, REESINK, Ger, FOLEY, Robert A. and LEVINSON, Stephen C. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science*, 309.2072-75.

FOLEY, William A. 1998. Toward understanding Papuan languages. *Perspectives on the Bird's Head*, ed. by Jelle Miedema, Cecilia Odé and Rien A. C. Dam, 503-18. Amsterdam: Rodopi.

FOLEY, William A. 2000. The languages of New Guinea. *Annual Review of Anthropology*, 29.357-404.

FRIEDLAENDER, Jonathan Scott. (ed.) 2007. *Genes, language, and culture history in the Southwest Pacific*. Oxford: Oxford University Press.

HUDJASHOV, Georgi and Toomas KIVISILD, Peter A. UNDERHILL, Phillip ENDICOTT, Juan J. SANCHEZ, Alice A. LIN, Peidong SHEN, Peter OEFNER, Colin RENFREW, Richard VILLEMS, Peter FORSTER. 2007. Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. *PNAS* Vol.104, no.21:8726-8730.

NICHOLS, Johanna. 1996. The comparative method as heuristic. *The comparative method revisited: Regularity and irregularity in language change*, ed. by Mark Durie and Malcolm Ross, 39-71. Oxford: Oxford University Press.

PAWLEY, Andrew K. 2007. Recent research on the historical relationships of the Papuan languages, or, What does linguistics say about the prehistory of Melanesia. *Genes, language, and culture history in the Southwest Pacific,* ed. by Jonathan Scott Friedlaender, 36- 58. New York: Oxford University Press.

REESINK, Ger. 2005. West Papuan languages: roots and development. *Papuan Pasts: cultural, linguistic and biological histories of Papuan-speaking peoples*, ed. by Andrew Pawley, Robert Attenborough, Jack Golson and Robin Hide, 185-218. Canberra: Pacific Linguistics.

ROSENBERG, Noah A., PRITCHARD, Jonathan K., WEBER, James L., CANN, Howard M., KIDD, Kenneth K., ZHIVOTOVSKY, Lev A. and FELDMAN, Marcus W. 2002. Genetic structure of human populations. *Science* 298:2381-2385.

ROSS, Malcolm. 2005. Pronouns as a preliminary diagnostic for grouping Papuan languages. *Papuan Pasts: cultural, linguistic and biological histories of Papuan-speaking peoples*, ed. by Andrew Pawley, Robert Attenborough, Jack Golson and Robin Hide, 15-65. Canberra: Pacific Linguistics.

SUMMERHAYES, Glenn. R. 2007. Island Melanseian Pasts: A view from archaeology. *Genes, language, and culture history in the Southwest Pacific,* ed. by Jonathan Scott Friedlaender, 10-35. New York: Oxford University Press.

WURM, S. A., LAYCOCK, D. C., VOORHOEVE, C. L., DUTTON, T. E. 1975. Papuan linguistic prehistory, and past language migrations in the New Guinea area. *New Guinea area languages and language study*, ed. by S. A. Wurm, 935-60. Canberra: Pacific Linguistics.